Week 7: The VC Dimension

Table of Contents

Definition of VC Dimension

1. Recap on growth function
   - When there exists break point k for an hypothesis set $\mathcal{H}$, growth function $m_{\mathcal{H}}(N)$ is bounded by:

$$m_{\mathcal{H}}(N) \leq B(N,k) = \sum_{i=0}^{k-1} \binom{N}{i}$$

     - Highest term from the combinatorics is $Nk - 1$

	$B(N,k)$	k = 1	2	3	4	5
	1	1	2	2	2	2
	2	1	3	4	4	4
	3	1	4	7	8	8
N	4	1	5	11	15	16
	5	1	6	16	26	31
	6	1	7	22	42	57

N^{k-1}	k = 1	2	3	4	5
1	1	1	1	1	1
2	1	2	4	8	16
3	1	3	9	27	81
4	1	4	16	64	256
5	1	5	25	125	625
6	1	6	36	216	1296

   - From the tables above, we can see that **provably and loosely**, for $N \geq 2, k \geq 3$

$$m_{\mathcal{H}}(N) \leq B(N,k) = \sum_{i=0}^{k-1} \binom{N}{i} \leq N^{k-1}$$

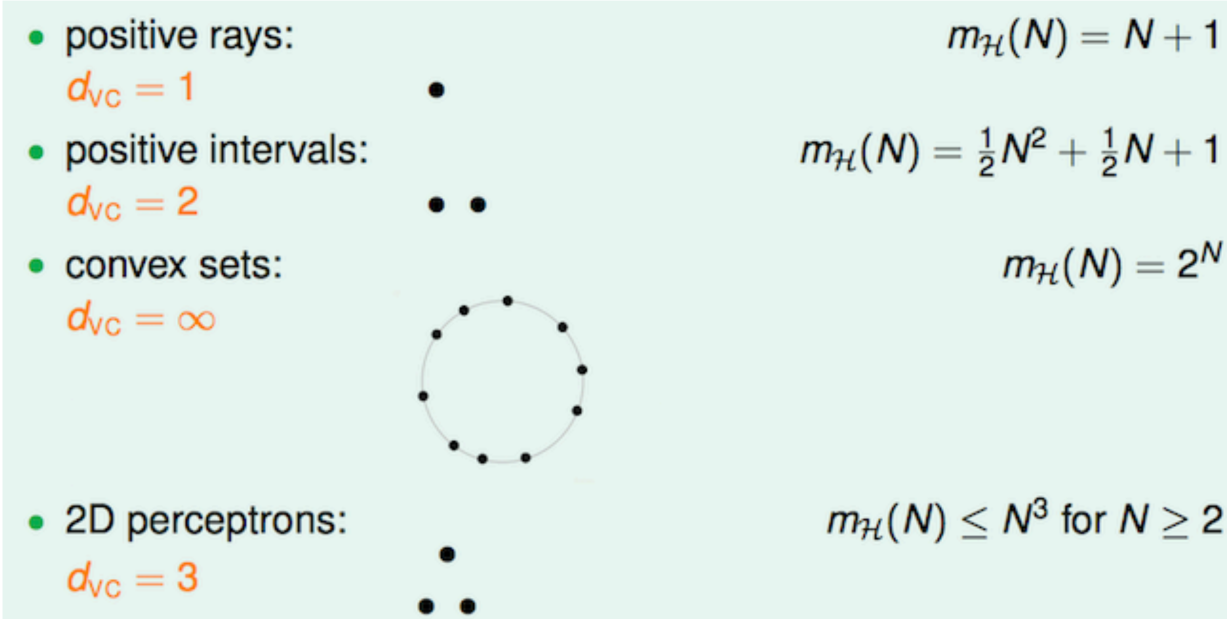
   - Plug the inequality above into Vapnik-Chervonenkis (VC) Bound gives:

For any $g = \mathcal{A}(\mathcal{D}) \in \mathcal{H}$ and 'statistical' large $\mathcal{D}$, for $\cancel{N \geq 2,}\ k \geq 3$

$$
\begin{aligned}
& \mathbb{P}_{\mathcal{D}}\Big[\big|E_{\text{in}}(g) - E_{\text{out}}(g)\big| > \epsilon\Big] \\
\leq \quad & \mathbb{P}_{\mathcal{D}}\Big[\exists h \in \mathcal{H} \text{ s.t. } \big|E_{\text{in}}(h) - E_{\text{out}}(h)\big| > \epsilon\Big] \\
\leq \quad & 4m_{\mathcal{H}}(2N)\exp\left(-\tfrac{1}{8}\epsilon^2 N\right) \\
\overset{\text{if } k \text{ exists}}{\leq} \quad & 4(2N)^{k-1}\exp\left(-\tfrac{1}{8}\epsilon^2 N\right)
\end{aligned}
$$

- In other words, learning is possible when:
  - The hypothesis set $\mathcal{H}$ has a break point k by which growth function $m_{\mathcal{H}}(N)$ is bounded
  - Sample size N is large enough to generalize $E_{out} \approxeq E_{in}$
  - Learning algorithm $\mathcal{A}$ is capable of picking an *optimal* hypothesis g with small E_{in}

2. VC Dimension
   - The formal name of **maximum non-** break point
   - In the context of 2D perceptron, VC dimension of $\mathcal{H}$, denoted $d_{VC}(\mathcal{H})$ is the **largest** N for which $m_{\mathcal{H}}(N) = 2^N$
     - In other words, d_{VC} is the **most** number of inputs that can be shattered by any hypothesis $h \in \mathcal{H}$
     - $d_{VC} = \min(k) - 1$
3. Implications of VC dimension
   - $N \leq d_{VC} \quad \Rightarrow \quad \mathcal{H}$ can shatter **some** N inputs
     - Note that **the reverse is not true**. Having a set of N inputs that cannot be shattered by $\mathcal{H}$ does not imply anything between $d_{VC}(\mathcal{H})$ and N (based on that information alone)
       - $\mathcal{H}$ might be able to shatter different input set from the same population, which would mean $d_{VC} \geq N$
       - Also possible that no input set of N from the population can be shattered by $\mathcal{H}$, in which case $d_{VC} < N$
   - $k > d_{VC} \quad \Rightarrow \quad k$ is a break point for $\mathcal{H}$
   - Combining with growth function above, *if* $N \geq 2, d_{vc} \geq 2,\ m_{\mathcal{H}}(N) \leq N^{d_{vc}}$



VC dimension and learning

1. The **worst case** generalization $E_{out}(g) \approx E_{in}(g)$ guaranteed by VC dimension is
   - Independent of learning algorithm $\mathcal{A}$
   - Independent of input distribution P
   - Independent of target function f
   - Available so long as sample size is large enough and break point exists

Generalizing PLA beyond 2D

1. VC dimension of n-D perceptron**: $d_{vc} = d + 1$

2. Proof Part 1: $d_{vc} \geq d + 1$

   - Recall that $d_{vc} \geq d + 1$ as long as there is **at least one** set of $d + 1$ inputs that can be shattered.

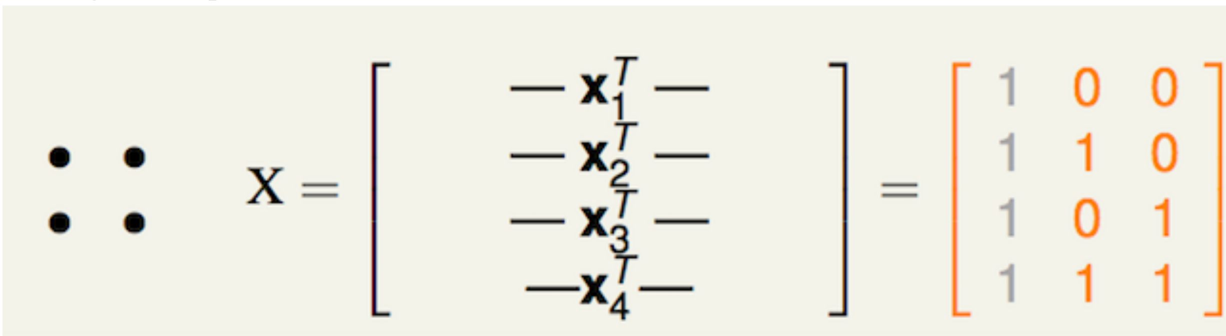
   - Given a set of trivial, **intervible** inputs $\mathbf{X}$

$$\mathbf{X} = \begin{bmatrix} -\mathbf{x}_1^T- \\ -\mathbf{x}_2^T- \\ \vdots \\ -\mathbf{x}_{d+1}^T- \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & & \ddots & 0 \\ 1 & 0 & \dots & 0 & 1 \end{bmatrix} \text{invertible}$$

   - Since the input matrix $\mathbb{X}$ is **invertible**, a weight vector w can be found such that:

- Given $y = \begin{bmatrix} y1 \\ . \\ . \\ . \\ y_{d+1} \end{bmatrix}$
- $sign(\mathbf{X}w) = y \Longleftarrow$ in special case $(\mathbf{X}w) = y \Longleftrightarrow w = \mathbf{X}^{-1}y$

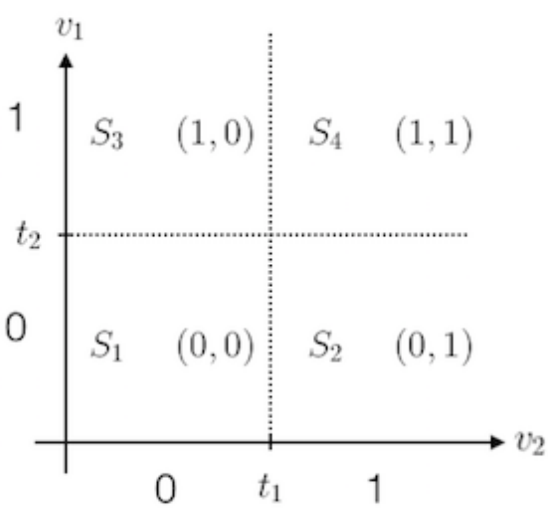
3. Proof Part 2: $d_{vc} \leq d + 1$
   - Recall that $d_{vc} \leq d + 1$ is guaranteed only when we can proof that it is impossible to shatter **any** set of $d + 2$ inputs
   - Starting from a special case in 2D



   - Given that there's now **more rows than columns** in the input matrix (d-dimension, 1 constant, d+2 data points), there must be some kind of **linear dependency** among the points
     - For the input set above, there exists linear dependency:

$$w^T x_4 = w^T x_2 + w^T x_4 - w^T x_1$$

     - If we map each of the points by their (x, y) into four quarnts in 2D, as shown below. We can see that the weight vector w must be **positive** for x_2, x_3, and **negative** for x_1 (or vice versa, so long as we treat all points with at least one non-zero axis to be the same sign).



     - The linear dependency above **mandates** that the weight(sign) for x_4 can **only be positive**

$$\mathbf{w}^T\mathbf{x}_4 = \underbrace{\mathbf{w}^T\mathbf{x}_2}_{\circ} + \underbrace{\mathbf{w}^T\mathbf{x}_3}_{\circ} - \underbrace{\mathbf{w}^T\mathbf{x}_1}_{\times} > 0$$

    - Linear dependency **restricts dichotomy**

  - Generalizing the special case to n-dimensional

    - Same linear dependency exists among points in d-D, when there are d-dimensions, 1 constant, and $n+2$ points in the input set

$$\mathrm{X} = \begin{bmatrix} -\mathbf{x}_1^T- \\ -\mathbf{x}_2^T- \\ \vdots \\ -\mathbf{x}_{d+1}^T- \\ -\mathbf{x}_{d+2}^T- \end{bmatrix}$$

more rows than columns:

linear dependence (some a_i non-zero)

$$\mathbf{x}_{d+2} = a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \ldots + a_{d+1}\mathbf{x}_{d+1}$$

    - Similar to the previous case, it is possible to represent x_{n+1} (and one step further, the product $w^T x_{n+1}$) as a **sum** of products between all the other points and their respective weight (some positive, some negative, some can be zero)

    - The linear dependency again mandates that $w^T x_{d+2}$ can only be positive (or negative if we view the signs another way), making some dichotomies impossible. The input set therefore cannot be shattered

$$\mathbf{w}^T\mathbf{x}_{d+2} = a_1\underbrace{\mathbf{w}^T\mathbf{x}_1}_{\circ} + a_2\underbrace{\mathbf{w}^T\mathbf{x}_2}_{\times} + \ldots + a_{d+1}\underbrace{\mathbf{w}^T\mathbf{x}_{d+1}}_{\times} > 0(\text{contradition!})$$

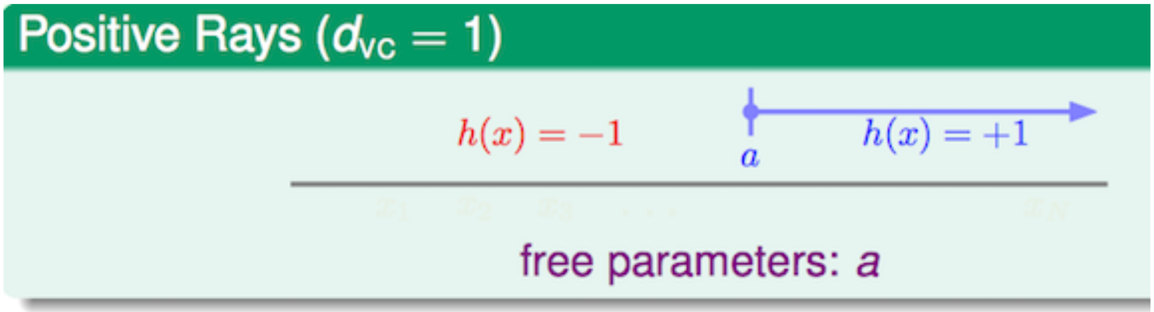
    - Since such linear depedency can be found for **any** input set of $d+2$ points in d-dimension, $d_{vc} \leq d+1$

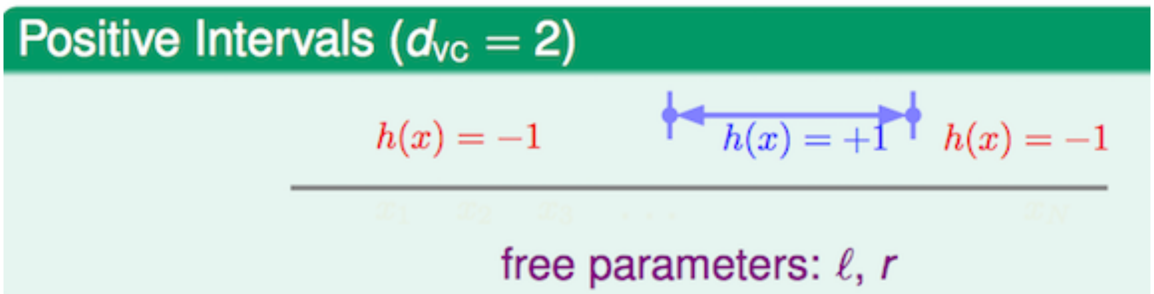
4. Combining the two-part proofs above result in the generalizable conclusion $d_{vc} = d+1$ for d-dimensional perceptron

Degrees of Freedom

1. Hypothesis parameters $w = (w_0, w_1, \ldots, w_d)$ **creates degrees of freeedom**
   - Different sets of w results in different hypotheses
2. Degrees of freedom:
   - **Analog** degress of freedom: Measured by hypothesis *quantity* $M = |\mathcal{H}|$
     - Shear number of possible hypothesis in a set, by introducing different sets of parameters
     - Depending on the hypothesis set, many hypothesis might belong to the same dichotomy
   - **Effective 'binary'** degrees of freedom: Measured by VC dimension $d_{vc} = d+1$

        - Number of parameters that could affect the number of dichotomies produced
            - Soem parameters are present and tunable, but do not lead to different dichotomies regardless of value chosen
        - $d_{vc}(\mathcal{H})$ is the *powerfulnness* of hypothesis set $\mathcal{H}$
        - $d_{vc} \approx$ # free parameters (not always)





3. VC dimension and objectives of learning
    - Recall the objectives of learning are:

        i. Make $E_{out}(g)$ close enough to $E_{in}(g)$ such that the learned model remains effective on unseen data

        ii. Make $E_{in}(g)$ small enough such that the model is a good estimation of the target function

    - Recall that finite bin Hoeffding's Inequalit guarantees that the probability of encountering a bad sample, for which E_{in} and E_{out} are very different, is bounded by $4 \cdot (2N)^{d_{vc}} \cdot exp(\ldots)$
        - **Small VC dimension**
            - Small chance of encountering bad sample. Good for learning objective 1
            - Small degree of freedom (choice), might not be possible to learn a good model from training set. Bad of objective 2.
        - **Large VC dimension**
            - Large chance of encountering bad sample. Bad for learning objective 2
            - Large degree of freedom, have many candidate models to choose from. Good for objective 2.
        - Important to choose the *right* d_{vc} to balance the tradeoff between learning objectives

Interpreting VC Dimension

1. Model complexity
   Per finite bin Hoeffding's with VC dimension, for any $g = \mathcal{A}(\mathcal{D}) \in \mathcal{H}$ and **statistically large** $\mathcal{D}$, if $d_{vc} \geq 2$

$$\mathbb{P}[BAD] = \mathbb{P}_{\mathcal{D}}[E_{in}(h) - E_{out}(h)| > \epsilon] \leq 4(2N)^{d_{vc}} exp(-\frac{1}{8}\epsilon^2 N)$$

$$\text{set} \quad \delta = 4(2N)^{d_{vc}} exp(-\frac{1}{8}\epsilon^2 N)$$

$$\text{then} \quad \epsilon = \sqrt{\frac{8}{N} ln(\frac{4(2N)^{d_{vc}}}{\delta})}$$

Generalization error, $|E_{in}(g) - E_{out}(g)|$ thus satisfies

$$E_{in}(g) - \sqrt{\frac{8}{N} ln(\frac{4(2N)^{d_{vc}}}{\delta})} \leq E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} ln(\frac{4(2N)^{d_{vc}}}{\delta})}$$

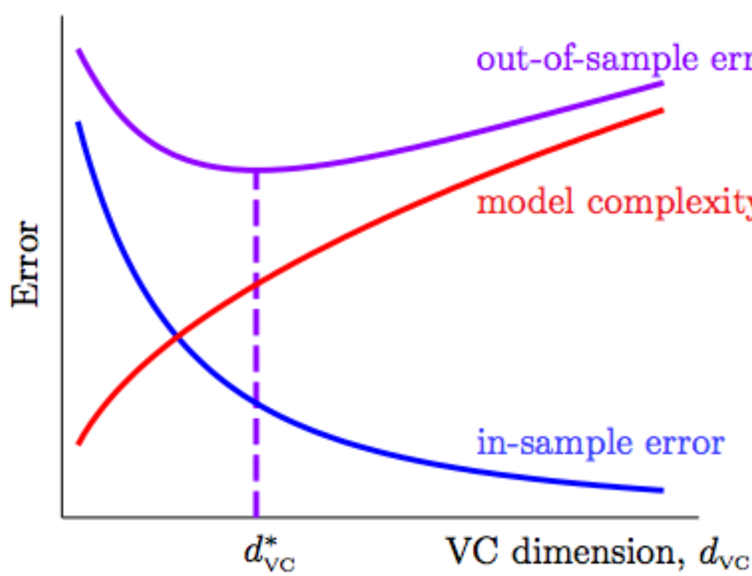
The **upper bound** here is of our main interest, as it determines how well would a model learned from input data would perform when used on other sample drawn from the same underlying distribution.

Penalty for model complexity is often denoted as:

$$\Omega(N, \mathcal{H}, \delta) = \sqrt{\frac{8}{N} ln(\frac{4(2N)^{d_{vc}}}{\delta})}$$

With a high probability:

$E_{out}(g) \leq E_{in}(g) + \Omega(N, \mathcal{H}, \delta)$
$\Rightarrow$ Complex model can potentially lead to large difference between E_{in} and E_{out}
$\Rightarrow$ "Overfitted"



2. Sample complexity

- The loose nature of VC bound often lead to dramatic difference between theoretical sample complexity and the practical value, given a error tolerance given **specs** $\epsilon = 0.1$, $\delta = 0.1$, $d_{VC} = 3$, want $4(2N)^{d_{VC}} \exp\left(-\frac{1}{8}\epsilon^2 N\right) \leq \delta$

N	bound
100	2.82×10^7
1,000	9.17×10^9
10,000	1.19×10^8
100,000	1.65×10^{-38}
29,300	9.99×10^{-2}

sample complexity:
need $N \approx 10,000 d_{VC}$ in theory

- Practical rule of thumb: $N \approx 10 d_{vc}$ is **often enough**

3. Looseness of VC Bound

- The significant difference bewteen theoretical and practical model complexity illustrates the looseness of VC Bound. This looseness is the result of following factors combined:

  a. Using Hoeffding's Inequality for **unknown** E_{out}
     - This allows VC Bound to be valid for any target distribution
  b. Using growth function $m_{\mathcal{H}}$ intead of a specific sample $|\mathcal{H}(x_1, \ldots, x_N)|$
     - This allows use of any sample from the population
  c. Using the **upper bound** of growth function $N^{d_{vc}}$, instead of growth function $m_{\mathcal{H}}(N)$ of a specfic hypothesis set $\mathcal{H}$
     - This allows use of **any hypothesis set** $\mathcal{H}$ with the same (readily specified) d_{vc}
  d. Using union bound on worst cases
     - This allows VC Bound to be valid regardless of choice made by learning algorithm $\mathcal{A}$

- Despite its looseness, it's hard to find stricter bound with the same guarantees.

- Futhermore, VC Bound holds *similar looseness for all models*, so it can still be used to compare model performance for the purpose of model selection