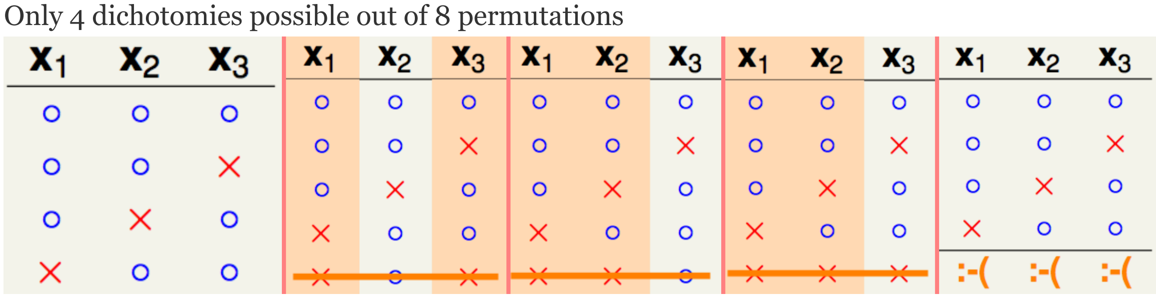
Week 6: Theory of Generalization

Table of Contents

Restriction of Break Point

1. General idea: The existence of break point k can dramatically reduce/limit the possible number of dichotomies. In other words, the max range of growth function $m_{\mathcal{H}}(N)$, and in turn, the value of M in finite bin Hoeffding's is greatly reduced when there exists a break point k for certain hypothesis set.
2. By example: What *must be true* when minimum break point $k = 2$
    - $N = 1$: every $m_{\mathcal{H}}(N) = 2$ by definition
    - $N = 2$: every $m_{\mathcal{H}}(N) < 4$ by definition $\rightarrow$ maximum possible 3 dichotomies
    - Max possible $m_{\mathcal{H}}(N)$ when $N = 3, k = 2$
        - Cannot shatter any 2 (out of 3) points by definition of break point
        - Only 4 dichotomies possible out of 8 permutations



3. Theory:

$$m_{\mathcal{H}}(N) \leq \text{max possible } \mathcal{H}(N) \text{ given } k \leq \textit{polynomial}(N)$$

Bounding Function: Basic Cases

1. Definion: **Bounding function** $B(N, k)$ is the maximum possible $m_{\mathcal{H}}(N)$ when break point = k
    - Combinatorial quantity: Max number of length-N vectors with (o, x), while **'no shatter' on any length-k** subvectors
    - Irrelevant of details of $\mathcal{H}$

- e.g. $B(N, 3)$ bounds both positive intervals (k=3) and 1D perceptrons (k=3)
- Upper bound of growth function is solely related to value of break point

2. Table of Bounding Function

$B(N,k)$	1	2	3	4	5	6	...
1							
2		3					
3		4					
4							
5							
6							
⋮							

$B(N,k)$	1	2	3	4	5	6	...
1	1						
2	1	3					
3	1	4					
4	1						
5	1						
6	1						
⋮	⋮						

$B(N,k)$	1	2	3	4	5	6	...
1	1	2	2	2	2	2	...
2	1	3	4	4	4	4	...
3	1	4		8	8	8	...
4	1				16	16	...
5	1					32	...
6	1						...
⋮	⋮						

$B(N,k)$	1	2	3	4	5	6	...
1	1	2	2	2	2	2	...
2	1	3	4	4	4	4	...
3	1	4	7	8	8	8	...
4	1			15	16	16	...
5	1				31	32	...
6	1					63	...
⋮	⋮						⋱

  - $B(2, 2) = 3$ (because given 2 points, cannot shatter → dichotomies < 4)
  - $B(3, 2) = 4$ (see previous section)
  - $B(N, 1) = 1$ (cannot shatter any one point, only one possible dichotomy regardless of the sign chosen for each individual points)
  - $B(N, k) = 2^N for N < k$ (number of points smaller than minimum break point → all 2^N dichotomies are possible)
  - $B(N, k) = 2^N - 1 for N = k$ (removing one dichotomy in order to satisfy the "breaking condition")

Bounding Functiion: Inductive Cases

1. Estimating value of $B(4, 3)$

  - In order to continue down the table of bounding function and extend it to more data points and higher break points, it's our interest to find the inherent relationship between $B(N, K)$ at higher N, k and lower N, k
  - The 11 dichotomies bounded by $B(4, 3)$ can be grouped as follows. Note that dichotomies in orange appear in *pairs*, with $x4$ alternating, whereas dichotomies in purple appear singled, with only one value possible for $x4$ without breaking the shattering constraint.

	x_1	x_2	x_3	x_4
2α	○	○	○	○
	○	○	○	×
	×	○	○	○
	×	○	○	×
	○	×	○	○
	○	×	○	×
	○	○	×	○
	○	○	×	×
β	×	×	○	×
	×	○	×	○
	○	×	×	○

- Ignoring $x4$ from table above (and reducing dichotomies down to $\alpha + \beta$) results in the dichotomies bounded by $B(3,3) \Rightarrow \alpha + \beta \leq B(3,3)$
  - Because when the scope shrinks down to 3 data points, the valid dichotomies must guarantee no shattering across **all 3**, in order to guarranttee no shattering across **any 3 out of 4** points when $x4$ is introduced.

	x_1	x_2	x_3
α	○	○	○
	×	○	○
	○	×	○
	○	○	×
β	×	×	○
	×	○	×
	○	×	×

- In addition, in order for dichotomies in orange to come in valid *pairs* when bounded by $B(4,3)$, they **must not shatter any 2 points**, such that values of $x4$ can be introduced in pairs. $\Rightarrow \alpha \leq B(3.2)$
- Combining constraints found above:

$$\begin{aligned} B(4,3) &= 2\alpha + \beta \\ \alpha + \beta &\leq B(3,3) \\ \alpha &\leq B(3,2) \\ \Rightarrow B(4,3) &\leq B(3,3) + B(3,2) \end{aligned}$$

2. Generalizing via induction

   - Provides **upper bound** of bounding function

$$
\begin{aligned}
B(N,k) &= 2\alpha + \beta \\
\alpha + \beta &\le B(N-1,k) \\
\alpha &\le B(N-1,k-1) \\
\Rightarrow B(N,k) &\le B(N-1,k) + B(N-1,k-1)
\end{aligned}
$$

		k					
	$B(N,k)$	1	2	3	4	5	6
	1	1	2	2	2	2	2
	2	1	3	4	4	4	4
	3	1	4	7	8	8	8
N	4	1	≤ 5	11	15	16	16
	5	1	≤ 6	≤ 16	≤ 26	31	32
	6	1	≤ 7	≤ 22	≤ 42	≤ 57	63

Bounding Function: The theorem

1. Using boundary and inductive formula, we can derive the following upper bound of bounding function $B(N,k)$:
   - Key is that for *fixed break point k*, $B(N,k)$ is upper bounded by $polynomial(N) \Rightarrow m_{\mathcal{H}}(N)$ is $polynomial(N)$if **break point exists**
   - Growth function $m_{\mathcal{H}}(N)$ is now bounded by one break point

$$B(N,k) \le \underbrace{\sum_{i=0}^{k-1} \binom{N}{i}}_{\text{highest term } N^{k-1}}$$

   - **Note**: Actually, the $\le$ in equation above **always** come out as equal

Introducing VC Bound

Revisiting Probability of Bad Samples

1. Given input set N, the probability for any hypothesis h in hypothesis set $\mathcal{H}$, and end up having very different out-of-sample error $E_out(h)$ compared to in-sample error $E_in(h)$, is bounded by finite-bin Hoeffinding's Inequality and growth function as:

$$\mathbb{P}[\exists h \in \mathcal{H} \text{ s.t. } |E_{in}(g) - E_{out}(g) > \epsilon|] \le 2 \cdot m_{\mathcal{H}}(N) \cdot exp(-2\epsilon^2 N)$$

2. The inequality above bascially replaced the number of hypothesis M in finite-bin Hoeffinding with growth

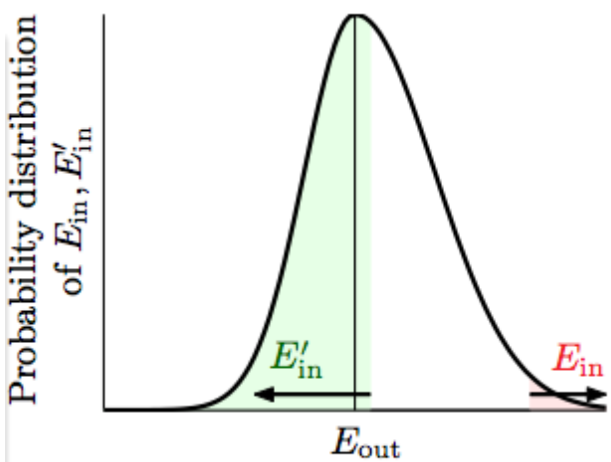
function $m_{\mathcal{H}}(N)$. Since we have proved above that growth function is bounded whenever there exists a break point, the substitution is valid.

3. When **N is large enough**, the inequality above becomes:

$$\mathbb{P}\left[\exists h \in \mathcal{H} \text{ s.t. } \left|E_{\text{in}}(h) - E_{\text{out}}(h)\right| > \epsilon\right] \leq 2 \cdot 2 m_{\mathcal{H}}(2N) \cdot \exp\left(-2 \cdot \frac{1}{16}\epsilon^2 N\right)$$

Sketch Proof

1. Replace E_out by E_in'
   - The number of possible values of $E_{in}(h)$ is finite (= number of dichotomies). However, number of $E_{out}(h)$ is infinite
     - In 2D perceptron, each dichotomy covers a collection of **infinite** number of lines that produce the same set of predictions on input data.
     - Each of the lines will have a different E_{out}
   - Replace non-traceable E_{out} with traceable E'_{in}, calculated from a **verification set** $\mathcal{D}'$ (also called *ghost data*) sampled from the same population as training set $\mathcal{D}$
   - Since training set $\mathcal{D}$, verfication set $\mathcal{D}'$, and out-of-sample set $\mathcal{D}_{out}$ come from the same population, when sample size N is large enough, the errors E_{in}, E'_{in} follows a Gaussian distribution with expected value E_{out}.



   - Given the Gaussian distribution shown above, when $|E_in(h) - E_out(h)|$ is large, there is 50% chance that $|E_{in}(h) - E'_{in}(h)|$ is **same or larger**. Therefore we have:

$$\frac{1}{2}\mathbb{P}\left[\exists h \in \mathcal{H} \text{ s.t. } \left|E_{\text{in}}(h) - E_{\text{out}}(h)\right| > \epsilon\right]$$
$$\leq \quad \mathbb{P}\left[\exists h \in \mathcal{H} \text{ s.t. } \left|E_{\text{in}}(h) - E'_{\text{in}}(h)\right| > \frac{\epsilon}{2}\right]$$

2. Decompose $\mathcal{H}$
   - Introduction of verification set $\mathcal{D}'$ **doubles** the number of samples, $N \Rightarrow 2N$
     - $\mathcal{H}(x_1, \ldots, x_N) \Rightarrow \mathcal{H}(x_1, \ldots, x_N, x'_1, \ldots, x'_N)$
     - Given that training set $\mathcal{D}$ and verification set $\mathcal{D}'$ are **mutually exclusive**, union bound

applies. Growth function accounting for verification set becomes $m_{\mathcal{H}}(2N)$

- Probability of encountering bad sample therefore becomes bounded by:

$$\begin{aligned} \text{BAD} &\leq 2\mathbb{P}\left[\exists h \in \mathcal{H} \text{ s.t. } \left|E_{\text{in}}(h) - E'_{\text{in}}(h)\right| > \frac{\epsilon}{2}\right] \\ &\leq 2m_{\mathcal{H}}(2N)\mathbb{P}\left[\text{fixed } h \text{ s.t. } \left|E_{\text{in}}(h) - E'_{\text{in}}(h)\right| > \frac{\epsilon}{2}\right] \end{aligned}$$

3. Use Hoeffding without replacement
    - Imagine a combined sample of size $2N$, from which N samples are chosen as input set $\mathcal{D}$, leaving the rest as verification set $\mathcal{D}'$ (again, input $\mathcal{D}$ and $\mathcal{D}'$ are mutually exclusive)
    - Given a fixed hypothesis h, and its error on input and verification set, $E_{in}(h)$ and $E'_{in}(h)$, respectively, its error on the *combine* sample can be thought as the **average** of errors on individual sets. In other words,

$$E_{combined} = \frac{E_{in} + E'_{in}}{2}$$

    - Given that $\mathcal{D}$ and $\mathcal{D}'$ are sampled randomly **without replacement** from the combined set, and that $E_{in}, E'_{in} \sim Normal(E_{out})$, the error upper bound is further restricted (halved):

$$|E_{in} - E'_{in}| > \frac{\epsilon}{2} \iff |E_{in} = \frac{E_{in} + E'_{in}}{2}| > \frac{\epsilon}{4}$$

        - Smaller bin, smaller ϵ, **Hoeffding's Inequality without replacement**

$$\begin{aligned} \text{BAD} &\leq 2m_{\mathcal{H}}(2N)\mathbb{P}\left[\text{fixed } h \text{ s.t. } \left|E_{\text{in}}(h) - E'_{\text{in}}(h)\right| > \frac{\epsilon}{2}\right] \\ &\leq 2m_{\mathcal{H}}(2N) \cdot 2\exp\left(-2\left(\frac{\epsilon}{4}\right)^2 N\right) \end{aligned}$$

VC Bound

Expanding the inequality above results in **Vapnik-Chervonenkis Bound**, or **VC Bound**

$$\begin{aligned} &\mathbb{P}\left[\exists h \in \mathcal{H} \text{ s.t. } \left|E_{\text{in}}(h) - E_{\text{out}}(h)\right| > \epsilon\right] \\ \leq\ &4m_{\mathcal{H}}(2N)\exp\left(-\frac{1}{8}\epsilon^2 N\right) \end{aligned}$$