PAC Learning

Binary Classification

$H \equiv$ Hypothesis Class

$h \in H \quad \mathbb{R}^d \longrightarrow \{-1, +1\}$

Dichotomy Vector

$x_1, \dots, x_N \in \mathbb{R}^d$

$(h(x_1), \dots, h(x_N)) \in \{-1, +1\}^N$

Dichotomy Set

$H(x_1, \dots, x_N) = \{(h(x_1), \dots, h(x_N)), h \in H\}$

$|H(x_1, \dots, x_N)| \leq 2^N$

Linear Classification

$d = 2$

x_1, x_2, x_3 are non colinear

$|H(x_1, \dots, x_N)| \leq 2^3 = 8$

If x_1, x_2, x_3 are colinear

$|H(x_1, \dots, x_N)| \leq 6$

Growth Function

$H \equiv$ Hypothesis Class

$$m_H(N) = \max_{x_1, \dots, x_N} |H(x_1, \dots, x_N)|$$

Linear Classifiers $d = 2$

$N = 3$, non colinear, $|H(x_1, \dots, x_N)| = 8$

$N = 3$, colinear, $|H(x_1, \dots, x_N)| = 6$

$m_H(3) = 8$

$N = 4$, $m_H(4) = 14$

$$\Pr\left(\bigcup_n |E_{in}(h) - E_{out}(h)| > \varepsilon\right)$$

Break Point

Let k be an integer such that $m_H(k) < 2^k$. Then k is called a break point for H

Example, Linear Classifier $d = 2$

$m_H(3) = 8$, $m_H(4) = 14$

$k = 4$ is break point for H

VC Dimension (Vapnik Chervonenkis)

Let N be an integer such that

$$m_H(N) = 2^N$$

$$m_H(N+1) < 2^{N+1}$$

i.e. $(N+1)$ is a break point for H

The VC dimension of H

$$d_{VC}(H) = N$$

Example, Linear Classification, $d=2$, $d_{VC}(H) = 3$

Example,

Consider a hypothesis class H such that $m_H(1) = 2$, $m_H(2) = 3$, find the maximum value of $m_H(3)$

$$m_H(3) = 4$$

x_1	x_2	x_3
+	+	+
+	+	−
+	−	−
−	−	−

Let H be a hypothesis class with VC dimension

$$d = d_{VC}(H)$$

Then the growth function $m_H(N)$ is upper bounded as

$$m_H(N) \leq \binom{N}{0} + \binom{N}{1} + \dots + \binom{N}{d}$$

$$= \sum_{k=0}^{d} \binom{N}{k} \leq N^d + 1$$

Note

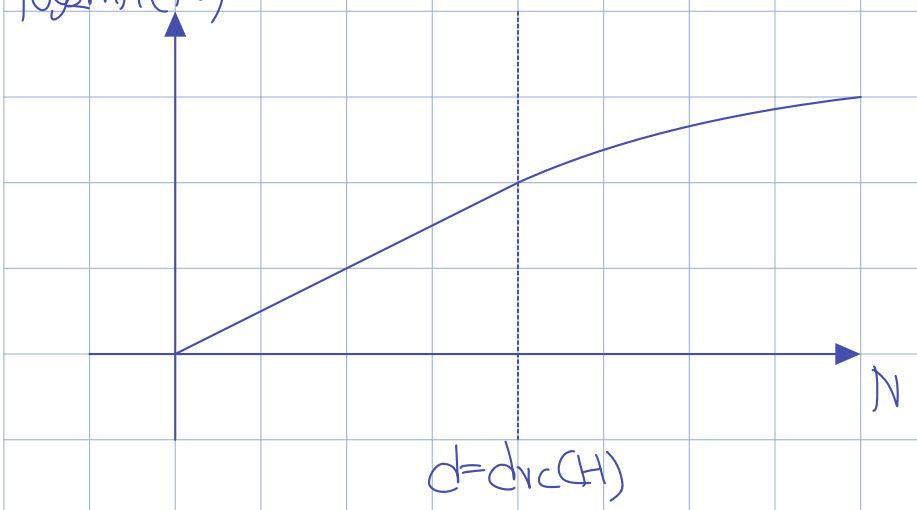
$N \leq d$

$m_H(N) = 2^N \qquad \log_2 m_H(N) = N$

$N > d$

$m_H(N) \leq N^d + 1 \qquad \log_2 m_H(N) \leq \log_2(N^d + 1)$

$\simeq d \log_2 N$



Proof of the upper bound (Textbook)(Chapter 2)

Main Theorem

$g \in H$ output hypothesis of learning algorithm

$$\Pr\left(\left| E_{in}(g) - E_{out}(g) \right| > \varepsilon \right)$$

$$\leq \Pr\left(\bigcup_{n}^{n} \left| E_{in}(h) - E_{out}(h) \right| > \varepsilon \right)$$

$$= \sum_{\substack{n: \text{distinct} \\ \text{Dichotomy Vector}}} \Pr\left(\left| E_{in}(h) - E_{out}(h) \right| > \varepsilon \right)$$

# of terms can be upper bounded by $m_H(N)$

$$\leq k_1 \, m_H(2N) \, e^{-k_2 N \varepsilon^2}$$

$k_1 = 4$

$k_2 = \frac{1}{8}$

Proof (Appendix)

Take Away,

With Probability $\geq 1 - \delta$

$$\left| E_{out}(g) - E_{in}(g) \right| < \varepsilon$$

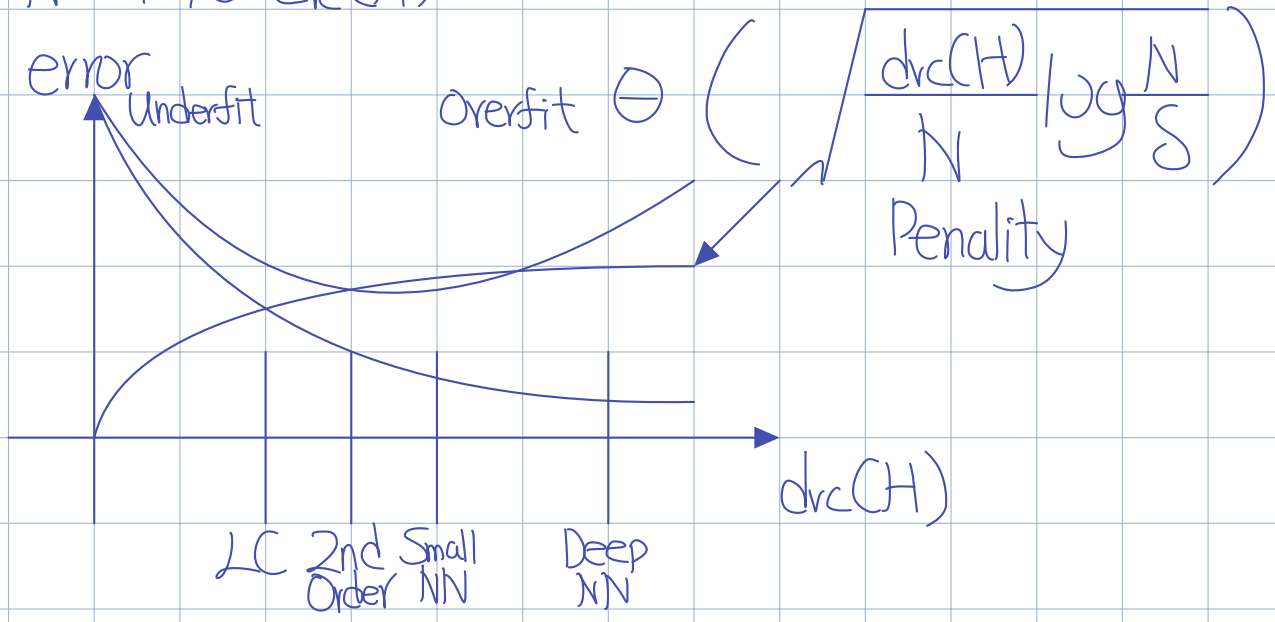
$$\delta = 4 m_H(2N) e^{-N\varepsilon^2/8}$$

$$\varepsilon = \sqrt{\frac{8}{N} \log \frac{4 m_H(2N)}{\delta}} = \Theta\left(\sqrt{\frac{d}{N} \log \frac{N}{\delta}} \right)$$

$$E_{out}(g) \leq E_{in}(g) + \Theta\left(\sqrt{\frac{d_{vc}(H)}{N}\log\frac{N}{\delta}}\right)$$

Rule Of Thumb

$N \gtrsim 10\, d_{vc}(H)$



VC Dimension Of Linear Classifier

Data Points $\underline{x} \in \mathbb{R}^{d+1}$

Classification $\hat{y} = \underline{\text{sign}(w^T x)}$

Rule $\underline{w} = (w_0, w_1, \dots, w_d) \in \mathbb{R}^{d+1}$

is weight vector

Claim $d_{vc}(H) = d + 1$

$d = 2$

$d_{vc}(H) = 3$

Part 1,

$N = d + 1$

Then there is a choice of $x_1, \dots, x_N$ that is shattered

Part 2,

$N = d + 2$ is a break point of H

Part 1,

Construct $x_1, \dots, x_N$ such that for any given label sequence $y_1, \dots, y_N \in \{-1, +1\}$ we will find a w such that

$$y_1 = \text{sign}(w^T x_1)$$
$$\vdots$$
$$y_N = \text{sign}(w^T x_N)$$

Ignore the sign function,

$$\left.\begin{aligned} y_1 &= w^T x_1 \\ &\vdots \\ y_N &= w^T x_N \end{aligned}\right\} \text{ Original Constraints Will be satisfied}$$

$$\begin{bmatrix} y_1 & \dots & y_N \end{bmatrix} = w^T \begin{bmatrix} x_1 & \dots & x_N \end{bmatrix}$$

$$x_i \in \mathbb{R}^{(d+1)\times 1}$$

$$Y = W^T X$$

$$X \in \mathbb{R}^{(d+1)\times(d+1)}$$

Choose any $x_1, \dots, x_N$ that are linearly independent

This is feasible as long as $N \leq d + 1$

$$\begin{matrix} 1 & & 1 \\ 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 1 \end{matrix}$$

Part 2,

$N = d + 2$

Then this set cannot be shattered by linear classifier

non invertible matrix

Can there be a choice of $x_1, \dots, x_N$ such that for any given $y_1, \dots, y_N \in \{-1, +1\}$, we can find some w that satisfy $y_i = \text{sign}(w^T x)$ for all i

Key Observation if $N > d+1$, $x_1, \dots, x_N \in \mathbb{R}^{d+1}$, they must be linearly independent

one of the vector, say x_1 must be linear combination of remaining ones

$x_1 = a_2 x_2 + \dots + a_N x_N$

for some choice $a_2, \dots, a_N \in \mathbb{R}$ not all zero

for $i = 2, 3, \dots, N$

let $y_i = \begin{cases} +1, & \text{if } a_i > 0 \\ -1, & \text{if } a_i < 0 \end{cases}$

Let $y_1 = -1$

Claim: There is no choice of w s.t.

$y_1 = \text{sign}(w^T x_1)$

$\vdots$

$y_N = \text{sign}(w^T x_N)$

Let w be s.t. $a_2 w^T x_2 \geq 0$

$y_2 = \text{sign}(w^T x_2)$ $\quad a_2 \geq 0, \; y_2 = +1, \; w^T x_2 > 0$

$\vdots$

$y_N = \text{sign}(w^T x_2)$ or $a_2 < 0, \; y_2 = -1, \; w^T x_2 < 0$

Hold for all

$x_1 \geq 0$

$w^T x_1 > 0$

$\text{sign}(w^T x_1) = 1$, $y_1 = -1$ cannot be achieved