HW - 6

Tanuoglua

Q1. Contd.

$N < k_0$ where k_0 is the smallest break point.

$\therefore \quad m_H(N) = 2^N$

$\Rightarrow \quad m_H(N)^2 = (2^N)^2 = 2^{2N}$

When $N < \frac{k_0}{2} \Rightarrow$ i.e., $2N < k_0$

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

When $N > \frac{k_0}{2} \Rightarrow$ i.e, $2N > k_0$

$$m_H(2N) < 2^{2N} = m_H(N)^2$$

$$\therefore \quad \boxed{m_H(2N) \leq m_H(N)^2}$$

Generalization Bound:

$$\begin{cases} m_H(N) \leq m_H\left(\frac{N}{2}\right)^2 \\ \text{or } m_H(N) \geq \sqrt{m_H(2N)} \end{cases} , \quad N \text{ is even}$$

$$m_H(N) \leq m_H\left(\frac{N+1}{2}\right)^2 , \quad N \text{ is odd}$$

Q1 Let $m_H(N) = k$

Now, if we partition any set of $2N$ points into two sets of N points each, each of these partition will produce at best k dichotomies. Now, combining the two sets, the maximum number of dichotomies possible will be the cross product of the two sets of dichotomies (with N points each)

i.e.,

$$m_H(2N) \leq k^2 = m_H(N)^2$$

$$\therefore \quad m_H(2N) \leq m_H(N)^2$$

Combining this result with the VC generalization bound, we get,

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

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

(Contd. on last page)

$m_H(N) = N+1$

Q2. (a) $N = 100$

Then, VC generalization bound,

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{100} \cdot \ln \frac{4(2 \cdot 100 + 1)}{0.1}}$$

$$\therefore E_{out}(g) \leq E_{in}(g) + 0.8481596$$

with probability at least 90%

$N = 10000$

Then, VC generalization bound,

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{10000} \cdot \ln \frac{4(2 \cdot 10000 + 1)}{0.1}}$$

$$\therefore E_{out}(g) \leq E_{in}(g) + 0.1042782$$

with probability at least 90%.

(b) $N = 100$, $d_{VC} = 1$

Then, VC generalization bound,

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} \ln \left(\frac{4((2N)^{d_{VC}} + 1)}{\delta} \right)}$$

$$\leq E_{in}(g) + \sqrt{\frac{8}{100} \ln \left(\frac{4 \cdot (200 + 1)}{0.1} \right)}$$

$$\leq E_{in}(g) + \sqrt{\frac{8}{100} \ln \frac{4 \cdot (200+1)}{0.1}}$$

$$\leq E_{in}(g) + 0.8481596$$

with probability at least 90%.

This is the same bound as previously seen in (a).

$N = 10000$

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{100} \ln \left(\frac{4 \cdot (2 \cdot 10000 + 1)}{0.1} \right)}$$

$$\leq E_{in}(g) + 0.1042782$$

with probability at least 90%.

This is the same bound as previously seen in (a).

(c). $d_{VC} = 10$ $\quad \delta = 0.05$ $\quad \epsilon = 0.05$

Then,

$$N \geq \frac{8}{(0.05)^2} \ln\left(\frac{4[(2N)^{10}+1]}{0.05}\right)$$

Let $N = 1000$ in RHS, then,

$$N \geq \frac{8}{(0.05)^2} \ln\left(\frac{4[(2\cdot 1000)^{10}+1]}{0.05}\right)$$
$$\approx 2.57251 \times 10^5$$

Substituting $N = 2.57251 \times 10^5$ in RHS, and continuing this process till convergence, we get,

$$N \approx 4.052957 \times 10^5$$

Q3.

(a) Let $d_{VC}(\mathcal{H}) = d$, then, we have, $m_{\mathcal{H}}(d) = 2^d$

So,

$$2^d = m_{\mathcal{H}}(d) = \max_{x_1,\cdots,x_d} |\mathcal{H}(x_1,\cdots,x_d)|$$
$$= \max_{x_1,\cdots,x_d} |\{(h(x_1),\cdots,h(x_d)) : h \in \mathcal{H}\}|$$
$$\leq |H| = M$$

So, we can say that $d \leq \log_2(M)$

so, basically,
if the VC dimension of a hypothesis set is d, then we need at least 2^d hypothesis to run the d points through in order to get them to shatter. So,

$$d \leq \log_2(M).$$

(b) At worst, $\bigcap_{k=1}^{K} \mathcal{H}_k = \{h\}$

In this case VC dimension = 0 as $m_{\mathcal{H}}(N) = 1 \quad \forall N$

So,

$$d_{VC}\left(\bigcap_{k=1}^{K} H_k\right) \geq 0$$

Now, let's assume,

$$d_{VC}\left(\bigcap_{k=1}^{K} \mathcal{H}_k\right) > \min_{1 \leq k \leq K} d_{VC}(\mathcal{H}_k) = d$$

which means that

$\bigcap_{k=1}^{K} \mathcal{H}_k$ can shatter $d+1$ points.

Let these points be $x_1, \cdots, x_{d+1}$.

Then,

$$\{-1, +1\}^{d+1} = \bigcap_{k=1}^{K} \mathcal{H}_k(x_1, \cdots, x_{d+1})$$

$$= \{(h(x_1), \cdots, h(x_{d+1})) : h \in \bigcap_{k=1}^{K} H_k\}$$

$$\subseteq \{(h(x_1), \cdots, h(x_{d+1})) : h \in H_k\}$$

$$= H_k(x_1, \cdots, x_{d+1}) \qquad \forall k = 1, \cdots, K$$

So,

$$2^{d+1} \leq |\{(h(x_1), \cdots, h(x_{d+1})) : h \in H_k\}| \leq 2^{d+1}$$

$$\Rightarrow |\{(h(x_1), \cdots, h(x_{d+1})) : h \in \mathcal{H}_k\}| = 2^{d+1}$$

$$\forall k = 1, \cdots, K$$

So, any $\mathcal{H}_k$ can shatter $d+1$ points

If we let $\min_{1 \leq k \leq K} d_{VC}(\mathcal{H}_k) = d_{VC}(\mathcal{H}_{k_0})$, we get,

$d = d_{VC}(\mathcal{H}_{k_0}) \geq d+1$ which is not possible

So,

$$\boxed{0 \leq d_{VC}\left(\bigcap_{k=1}^{K} \mathcal{H}_k\right) \leq \min_{1 \leq k \leq K} d_{VC}(\mathcal{H}_k)}$$

(c) Let $d_{VC}(\mathcal{H}_k) = d_k \quad \forall \quad k = 1, \cdots, K$

So, $\mathcal{H}_k$ shatters d_k points $x_1, \cdots, x_{d_k}$

$$\therefore \mathcal{H}_k(x_1, \cdots, x_{d_k}) = \{-1, +1\}^{d_k}$$
$$\forall \quad k = 1, \cdots, K$$

So,

$$\{-1, +1\}^{d_k} = \{(h(x_1), \cdots, h(x_{d_k})) : h \in \mathcal{H}_k\}$$
$$\subset \{(h(x_1), \cdots, h(x_{d_k})) : h \in \bigcup_{k=1}^{K} \mathcal{H}_k\}$$

So,

$$2^{d_k} \leq |\{(h(x_1), \cdots, h(x_{d_k})) : h \in \bigcup_{k=1}^{K} \mathcal{H}_k\}| \leq 2^{d_k}$$

So, we can say,

$$\Rightarrow |\{(h(x_1), \cdots, h(x_{d_k})) : h \in \bigcup_{k=1}^{K} \mathcal{H}_k\}| = 2^{d_k}$$

~~This is similar to proof done~~

$$\therefore m_{\bigcup_{k=1}^{K} H_k}(d_k) = 2^{d_k} \quad \forall \quad k = 1, \cdots, K$$

$$\Rightarrow d_{VC}\left(\bigcup_{k=1}^{K} H_k\right) \geq d_k \quad \forall \; k$$

$$\Rightarrow d_{VC}\left(\bigcup_{k=1}^{K} H_k\right) \geq \max_{1 \leq k \leq K} d_k = \max_{1 \leq k \leq K} d_{VC}(H_k)$$

- For upper bound, by induction proof:

Let $K = 2$ then, $d_{VC}(\mathcal{H}_1) = d_1$

$d_{VC}(\mathcal{H}_2) = d_2$

So, the number of dichotomies generated by $H_1 \cup H_2$ is the sum of the dichotomies generated by H_1 and H_2. So,

$$
\begin{aligned}
m_{H_1 \cup H_2}(N) &\leq m_{H_1}(N) + m_{H_2}(N) \\
&\leq \sum_{i=0}^{d_1} \binom{N}{i} + \sum_{i=0}^{d_2} \binom{N}{i} \\
&\leq \sum_{i=0}^{d_1} \binom{N}{i} + \sum_{i=0}^{d_2} \binom{N}{N-i} \\
&\leq \sum_{i=0}^{d_1} \binom{N}{i} + \sum_{i=N-d_2}^{N} \binom{N}{i} \\
&\leq \sum_{i=0}^{d_1} \binom{N}{i} + \sum_{i=d_1+1}^{N-d_2-1} \binom{N}{i} + \sum_{i=N-d_2}^{N} \binom{N}{i} \\
&= \sum_{i=0}^{N} \binom{N}{i} \\
&= 2^N \qquad \forall N
\end{aligned}
$$

s.t. $d_1 + 1 \leq N - d_2 - 1$

$\Rightarrow N \geq d_1 + d_2 + 2$.

So, we can say,

$$
\begin{aligned}
d_{VC}(H_1 \cup H_2) &\leq d_1 + d_2 + 1 \\
&\leq 1 + \sum_{k=1}^{2} d_{VC}(H_k)
\end{aligned}
$$

So, by induction, if this is correct for $K-1$, Then,

$$
d_{VC}\left(\bigcup_{k=1}^{K} H_k\right) = d_{VC}\left(\left(\bigcup_{k=1}^{K-1} H_k\right) \cup H_k\right)
$$

$$\therefore \quad d_{vc}\left(\bigcup_{k=1}^{K} \mathcal{H}_k\right) \leq 1 + d_{vc}\left(\bigcup_{k=1}^{K-1} \mathcal{H}_k\right) + d_{vc}(\mathcal{H}_K)$$

$$\leq 1 + (K-2) + \sum_{k=1}^{K-1} d_{vc}(\mathcal{H}_k) + d_{vc}(\mathcal{H}_K)$$

$$\boxed{\therefore d_{vc}\left(\bigcup_{k=1}^{K} \mathcal{H}_k\right) \leq K-1 + \sum_{k=1}^{K} d_{vc}(\mathcal{H}_k)}$$

So,

$$\boxed{\max_{1 \leq k \leq K} d_{vc}(\mathcal{H}_k) \leq d_{vc}\left(\bigcup_{k=1}^{K} \mathcal{H}_k\right) \leq K-1 + \sum_{k=1}^{K} d_{vc}(\mathcal{H}_k)}$$

Q5. Tikhonov regularization constraint

$$w^T \Gamma^T \Gamma w \leq C$$

(a) To obtain : $\sum_{q=0}^{Q} w_q^2 \leq C$

In order to obtain $w^T \Gamma^T \Gamma w = \sum_{q=0}^{Q} w_q^2$

we should use $\Gamma = I$

where I is identity matrix.

(b) To obtain : $\left(\sum_{q=0}^{Q} w_q\right)^2 \leq C$

In order to obtain $w^T \Gamma^T \Gamma w = \left(\sum_{q=0}^{Q} w_q\right)^2$

we should use $\Gamma = [1 \; 1 \; 1 \; 1 \; 1 \; \ldots \; 1]$

i.e., Γ should be a row of ones.

This implies that

$$w^T \Gamma^T = \sum_{q=0}^{Q} w_q$$

and thus,

$$w^T \Gamma^T \Gamma w = \left(\sum_{q=0}^{Q} w_q \right)^2$$

Q6. (a) $p(r) = \prod_{j=1}^{D} \text{Bur}\left(r_j \mid \Pi_0 \right)$

$$= \Pi_0^{\|r\|_0} \left(1 - \Pi_0 \right)^{D - \|r\|_0}$$

Here, $\Pi_0 = 0.2 = \frac{1}{5}$

$$D = 5$$

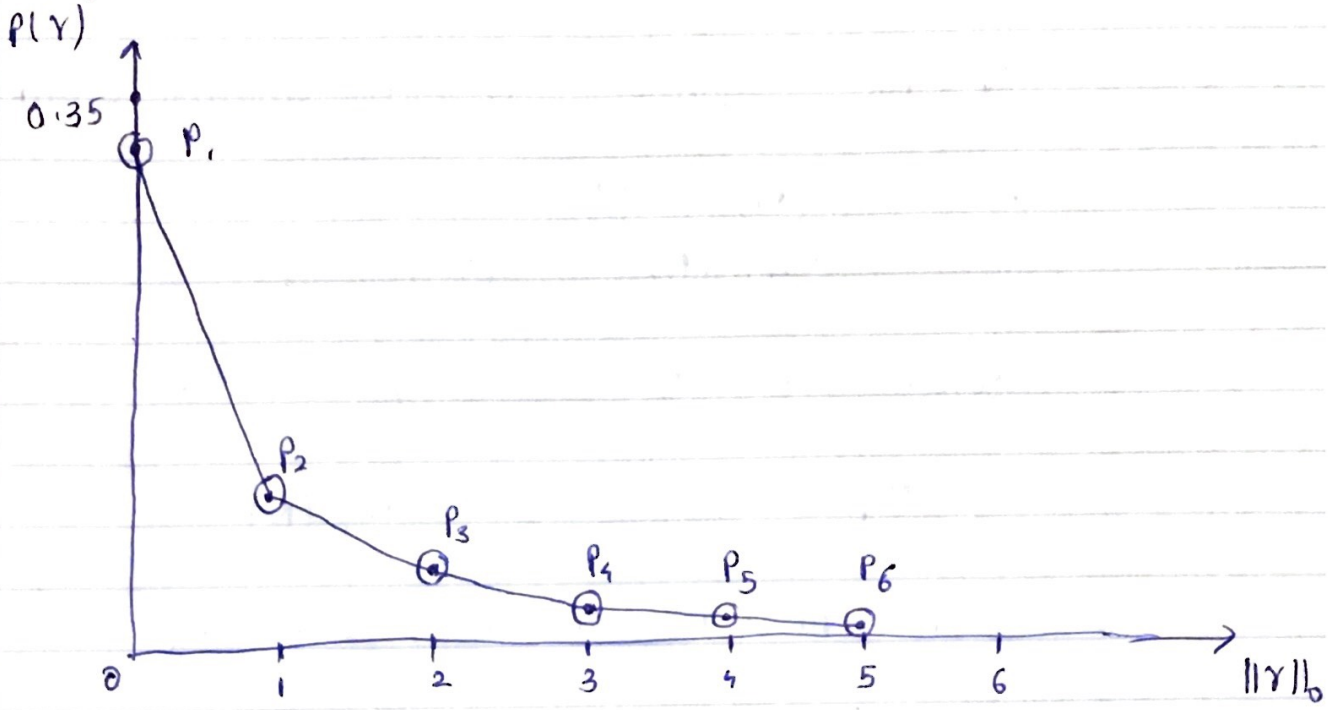
so,

$$p(r) = \left(\frac{1}{5}\right)^{\|r\|_0} \left(\frac{4}{5}\right)^{5 - \|r\|_0}$$

Substituting different values, we get,

$\|r\|_0$	$p(r)$	
0	0.32768	(P_1)
1	0.08192	(P_2)
2	0.02048	(P_3)
3	0.00512	(P_4)
4	0.00128	(P_5)
5	0.00032	(P_6)

So, plot :



(b) Adding a constraint that the norm of gamma should be greater than 0 would prevent the feature selection in this case from eliminating all features (i.e., resulting in $\gamma_i = 0 \ \forall i$).