HW 4

↳ Hardik Prajapati (2678294168)

1. $N = 4601$

$N_{Tr} = 3000$

$N_{Test} = 1601$

δ (tolerance) $= 0.1$

Given:- If H is a linear perceptron classifier in D Dimensions

$d_{VC}(H) = D+1$

Here, D = 57 features

a) $d_{VC}(H) = D+1 = 58$

b) $E_{out}(h_g) \leq E_{in}(h_g) + \varepsilon_{VC}$

$E_{in}(h_g) \rightarrow$ Training Data Set

$$\varepsilon_{VC} = \sqrt{\frac{8}{N} \ln \frac{4[(2N)^{d_{VC}} + 1]}{\delta}}$$

$$= \sqrt{\frac{8}{(3000)} \ln \frac{4[(2(3000))^{58} + 1]}{(0.1)}}$$

$$= 1.1642$$

c) $D = 10$

$N_{Tr} = 10{,}000$

$\Rightarrow d_{VC} = 11$

$$\varepsilon_{VC} = \sqrt{\frac{8}{(10{,}000)} \ln \frac{4\left[(2(10\,000))^{11} + 1\right]}{0.1}}$$

$$= 0.3001$$

d) $\varepsilon_{VC} = 0.1$

$\delta = 0.1$

$D = 10$

$d_{VC} = 11$

$$\varepsilon_{VC} = \sqrt{\frac{8}{N_{Tr}} \ln \frac{4\left[(2N_{Tr})^{d_{VC}} + 1\right]}{\delta}}$$

$$\Rightarrow 0.1 = \sqrt{\frac{8}{N_{Tr}} \ln \frac{4\left[(2N_{Tr})^{11} + 1\right]}{0.1}}$$

$$\Rightarrow N_{Tr} = 110000$$

e) $E_{in}(h_g) \rightarrow E_{test}(h_g)$

H = Hypothesis set = $\{h_g\}$

Cardinality = 1.

f) $E_{out}(h_g) \leq E_{test}(h_g) + \varepsilon$

$D = 57$

$N_{Test} = 1601$

$$\varepsilon_M = \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}}$$

$$\Rightarrow \varepsilon_M = \sqrt{\frac{1}{2(1601)} \ln \frac{2(1)}{0.1}}$$

$$= 0.0305$$

2. $X = \{1\} \times R^d$

To show:- $d_{vc} \geq d+1$ & $d_{vc} \leq d+1 \Rightarrow d_{vc} = d+1$

a) $x_0 = [0, 0, \cdots 0]^T_{1\times d}$

$x_1 = [1, 0, \cdots 0]^T_{1\times d}$

$x_2 = [0, 1, 0, \cdots 0]^T_{1\times d}$

$\vdots$

$x_d = [0, 0, \cdots, 1]^T_{1\times d}$

} $d+1$ points

Linear Perceptron Model :- $\text{sgn}[w^T x + w_0]$ $\{+1, -1\}$

Augmented column of 1

Now,

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & & \\ 1 & 0 & 0 & \cdots & 1 \end{bmatrix}_{(d+1)\times(d+1)}$$

Let $w' = \begin{bmatrix} w_0, w_1, w_2 \cdots w_d \end{bmatrix}_{(d+1)\times 1}$

$$h(\underline{\underline{x}}) = X w'$$

$$\begin{bmatrix} h(x_0) \\ h(x_1) \\ \vdots \\ h(x_d) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & & & & \\ 1 & 0 & 0 & \cdots & 1 \end{bmatrix}_{(d+1)\times(d+1)} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix}_{(d+1)\times 1}$$

$$\Rightarrow h(x) = \begin{bmatrix} w_0 \\ w_0 + w_1 \\ w_0 + w_2 \\ \vdots \\ w_0 + w_d \end{bmatrix}$$

Let, $w_0 = 0.5\, h(x_0)$

$$\Rightarrow \begin{bmatrix} h(x_0) \\ h(x_1) \\ \vdots \\ h(x_d) \end{bmatrix} = \text{Sgn} \begin{bmatrix} 0.5^0 h(x_0) \\ 0.5^1 h(x_0) + w_1 \\ \vdots \\ 0.5^d h(x_0) + w_d \end{bmatrix}$$

Now, If $w_i = h(x_i) \quad \forall\, i \in \{1, \dots d\}$

Then, Linear perceptron can classify all these $d+1$ points correctly for their different combinations.

$\Rightarrow$ The perceptron model can atleast shatter all the $d+1$ points.

$\Rightarrow d_{VC} \geq d+1$

b) Now, Let $x_{d+1} = [1, 0, 0, \dots 1]^T_{1 \times d}$

↑ Linear combination of x_1 & x_d.

$$\Rightarrow X = \begin{bmatrix} 1 & 0 & 0 \dots & 0 \\ 1 & 1 & 0 \dots & 0 \\ 1 & 0 & 1 \dots & 0 \\ & & & \\ 1 & 0 & 0 \dots & 1 \\ 1 & 1 & 0 \dots & 1 \end{bmatrix}_{(d+2) \times (d+1)}$$

Let $w_0 = 0.5 * h(x_0)$

$$\Rightarrow \begin{bmatrix} h(x_0) \\ h(x_1) \\ \vdots \\ \vdots \\ h(x_d) \\ h(x_{d+1}) \end{bmatrix} = \text{Sgn} \begin{bmatrix} 0.5 * h(x_0) \\ 0.5 * h(x_1) + w_1 \\ \vdots \\ \vdots \\ 0.5 * h(x_d) + w_d \\ 0.5 * h(x_0) + w_1 + w_d \end{bmatrix}$$

Again, $w_i = h(x_i) \quad \forall\, i \in \{1, \dots d\}$

Now,

i) Say, $h(x_0) = 1$, $h(x_1) = 1$, $h(x_d) = 1$

Assume $h(x_{d+1}) = 1$

$$\Rightarrow h(x_{d+1}) = \text{sgn}\left[(0.5)(1) + 1 + 1\right]$$

$$= 1$$

$\Rightarrow$ Correctly classified.

ii) Say, $h(x_0) = 1$, $h(x_1) = 1$, $h(x_d) = 1$

Assume $h(x_{d+1}) = -1$

$$\Rightarrow h(x_{d+1}) = \text{Sgn}\left[(0.5)(1) + 1 + 1\right]$$

$$= 1$$

$\Rightarrow$ Mis-classified.

$\Rightarrow$ The Linear perceptron cannot shatter $d+2$ points.

$\Rightarrow$ $d_{VC} \leq d+1$.

Ans:- From part (a) & part (b), it can be concluded that VC Dimension of Linear perceptron (with d+1 parameters, counting w_0) is exactly equal to d+1.

Hence, $d_{VC} = d+1$

3. AML Pb 2.24

$x \in [-1, 1]$

$D = \{\{x_1, x_1^2\}, \{x_2, x_2^2\}\}$

Hypothesis set :- $\{h(x) = ax+b \mid a \in R, b \in R\}$

a.i) $g^{(D)}(x)$ = Line fitting points $\{\{x_1, x_1^2\}, \{x_2, x_2^2\}\}$

Line :- $x_1^2 = a x_1 + b$

$x_2^2 = a x_2 + b$

$x_1^2 - x_2^2 = a\{x_1 - x_2\}$

$\Rightarrow a = \frac{x_1^2 - x_2^2}{x_1 - x_2}$

$\Rightarrow a = x_1 + x_2$, if $x_1 \neq x_2$

$$\Rightarrow b = x_1^2 - \{x_1 + x_2\} x_1$$

$$\Rightarrow b = x_1^2 - x_1^2 - x_1 x_2$$

$$\Rightarrow b = -x_1 x_2$$

$$\Rightarrow g^{(D)}(x) = (x_1 + x_2)x - x_1 x_2$$

a.2) $$\bar{g}(x) = E_D\left[(x_1 + x_2)x - x_1 x_2\right]$$

$$= E_D\left[x_1 x\right] + E_D\left[x_2 x\right] - E_D\left[x_1 x_2\right]$$

↳ Now, Here 'x' is not Random variable.

↳ Also, x_1 & x_2 are 2 independent Random variables, both having uniform Distribution $[-1, 1]$

Hence, $E_D[x_1] = E_D[x_2] = 0$

$$\Rightarrow \bar{g}(x) = x E_D[x_1] + x E_D[x_2] - E_D[x_1] E_D[x_2]$$

$$\Rightarrow \bar{g}(x) = 0$$

b) ↳ Fix $\underline{x}$

Repeat for 1000 times:
- ↳ Sample (Random) 2 points from uniform distribution $[-1, 1]$
- ↳ Fit $y = (x_1 + x_2)x + x_1 x_2$ & compute $g^D(\underline{x})$

↳ Compute average of $g^D(x)$, which will equal to $\bar{g}(x)$

Now, for out of sample error

Repeat for 5,000 times:
- ↳ Random Sample $\underline{x}$ from Unif $[-1, 1]$
- ↳ Compute $g^D(\underline{x})$ as shown above.
- ↳ We will have array for each $\underline{x}$. Take average of it, $\bar{g}(\underline{x})$
- ↳ Compute $E_D\left[\left(\bar{g}(\underline{x}) - g^D(\underline{x})\right)^2\right]$
- ↳ Compute $\left[\left(\bar{g}(\underline{x}) - f(\underline{x})\right)^2\right]$
- ↳ Compute $E_D\left[\left(g^D(\underline{x}) - f(\underline{x})\right)^2\right]$

↳ Take average & compute

$$E_x\left[E_D\left[\left(\bar{g}(\underline{x}) - g^D(\underline{x})\right)^2\right]\right] \rightarrow \text{Variance}$$

$$E_x\left[\left(\bar{g}(\underline{x}) - f(\underline{x})\right)^2\right] \rightarrow \text{Bias}$$

$$E_x\left[E_D\left[\left(g^{(D)}(\underline{x}) - f(\underline{x})\right)^2\right]\right] \rightarrow \text{Out of Sample Error}$$

c) → Simulation

d) From (a), $\bar{g}(\underline{x}) = 0$

$$\text{Var} = E_x\left[E_D\left[\left(g(\underline{x}) - g^D(\underline{x})\right)^2\right]\right]$$

$$= E_x\left[E_D\left[\left(0 - (x_1+x_2)\underline{x} + x_1x_2\right)^2\right]\right]$$

$$= E_x\left[E_D\left[(x_1+x_2)^2\underline{x}^2 + x_1^2x_2^2 - 2x_1x_2(x_1+x_2)\underline{x}\right]\right]$$

$$= E_D\left[E_x\left[(x_1+x_2)^2\underline{x}^2 + x_1^2x_2^2 - 2x_1x_2(x_1+x_2)\underline{x}\right]\right]$$

$$= E_D\left[(x_1+x_2)^2 E_x[\underline{x}^2] + x_1^2x_2^2 E_x[1] - 2x_1x_2(x_1+x_2)E_x[\underline{x}]\right]$$

Now,

$$E_x[\underline{x}] = 0 \quad , \quad E_x[\underline{x}^2] = \frac{1}{3}$$

$$E_D[x_1] = E_D[x_2] = 0 \quad , \quad E_D[x_1^2] = E_D[x_2^2] = \frac{1}{3}$$

$$\Rightarrow \text{Var} = E_D\left[(x_1+x_2)^2\left(\frac{1}{3}\right) + x_1^2x_2^2 - 0\right]$$

$$= \frac{1}{3}E_D[x_1^2 + x_2^2 + 2x_1x_2] + E_D[x_1^2x_2^2]$$

$$= \frac{1}{3}\left\{E_D[x_1^2] + E_D[x_2^2] + 2E_D[x_1]E_D[x_2]\right\} + E_D[x_1^2]E_D[x_2^2]$$

$$\Rightarrow Var = \frac{1}{3}\left\{\frac{1}{3} + \frac{1}{3} + 0\right\} + \left\{\frac{1}{3}\right\}\left\{\frac{1}{3}\right\}$$

$$= \frac{2}{9} + \frac{1}{9}$$

$$= \frac{3}{9}$$

$$= \frac{1}{3}$$

↳ $Bias = E_x\left[\left(\bar{g}(\underline{x}) - f(\underline{x})\right)^2\right]$

$$= E_x\left[\left(0 - \underline{x}^2\right)^2\right]$$

$$= E_x\left[\underline{x}^4\right]$$

$$= \frac{1}{5}$$

↳ $E_D\{E_{out}\} = E_x\left[E_D\left[\left(g^{(D)}(\underline{x}) - f(\underline{x})\right)^2\right]\right]$

$$= E_x\left[E_D\left[\left(\left(x_1 + x_2\right)\underline{x} - x_1 x_2 - \underline{x}^2\right)^2\right]\right]$$

$$\Rightarrow E_D\{E_{out}\} = E_x\Big[E_D\Big[\{(x_1+x_2)\bar{x} - x_1x_2\}^2 + \bar{x}^4 - 2\bar{x}\{(x_1+x_2)\bar{x} - x_1x_2\}\Big]\Big]$$

$$= E_x\Big[\underbrace{E_D\Big[\{(x_1+x_2)\bar{x} - x_1x_2\}^2\Big]}_{\frac{1}{3} = (Var)} + \underbrace{E_D[\bar{x}^4]}_{\bar{x}^4} - 2\bar{x}\,E_D\Big[(x_1+x_2)\bar{x} - x_1x_2\Big]\Big]$$

$$= E_x\Big[\frac{1}{3}\Big] + \underbrace{E_x[\bar{x}^4]}_{\frac{1}{5} = Bias} - 2\,E_D\Big[(x_1+x_2)\bar{x} - x_1x_2\Big]\underbrace{E_x[\bar{x}]}_{0}$$

$$\Rightarrow E_D\{E_{out}\} = \frac{1}{3} + \frac{1}{5}$$

$$= \frac{8}{15}$$

↳ $Var(\bar{x}) = E_D\Big[\big(g^{(D)}(x) - \bar{g}(x)\big)^2\Big]$

$$= E_D\Big[\big((x_1+x_2)\bar{x} + x_1x_2\big)^2\Big]$$

$$= E_D\Big[(x_1+x_2)^2\bar{x}^2 + x_1^2x_2^2 - 2x_1x_2(x_1+x_2)\bar{x}\Big]$$

$$\Rightarrow Var(\underline{x}) = \underline{x}^2 E_D\left[(x_1+x_2)^2\right] + E_D\left[x_1^2\right]E_D\left[x_2^2\right] - 2\underline{x}\, E_D\left[x_1^2 x_2 + x_2^2 x_1\right]$$

$$= \underline{x}^2\left\{E_D\left[x_1^2\right] + E_D\left[x_2^2\right] + 2E_D\left[x_1\right]E_D\left[x_2\right]\right\} + \left(\frac{1}{3}\right)\left(\frac{1}{3}\right) - 2\underline{x}\left\{E_D\left[x_1^2\right]E_D\left[x_2\right] + E_D\left[x_2^2\right]E_D\left[x_1\right]\right\}$$

$$= \underline{x}^2\left\{\frac{1}{3} + \frac{1}{3} + 0\right\} + \frac{1}{9} - 2\underline{x}\left\{0 + 0\right\}$$

$$= \underline{x}^2\left\{\frac{2}{3}\right\} + \frac{1}{9}$$

↳ $Bias(\underline{x}) = \left(\bar{g}(\underline{x}) - f(\underline{x})\right)^2$

$$= \left(0 - \underline{x}^2\right)^2$$

$$= \underline{x}^4$$

4. AML Pb 2.13 (a), (b)

(a) $H = \{h_1, h_2, \cdots h_M\}$

To prove $d_{vc}(H) \leq \log_2 M$

Now, $m_H(N) \leq N^{d_{vc}} + 1$

$\Rightarrow$ For, 'd_{vc}' no. of points, growth function is

$$m_H(d_{vc}) = 2^{d_{vc}} \quad \text{-(i)}$$

Let us assume that $d_{vc}(H) > \log_2 M$

$$\Rightarrow 2^{d_{vc}(H)} > M \quad \text{-(ii)}$$

$\Rightarrow$ From (i) & (ii),

$$m_H(d_{vc}) = 2^{d_{vc}} > M$$

$\Rightarrow$ No. of Hypothesis that shatter 'd_{vc}' no. of points is larger than M.

$\Rightarrow$ But, $H = \{h_1, h_2, \cdots h_M\}$ has only finite 'M' hypothesis.

$\Rightarrow$ Contradiction to our Assumption.

$\Rightarrow$ $d_{vc}(H) \leq \log_2 M$ //

Hence Proved

(b) Hypothesis Sets :- $H_1, H_2, \cdots H_k$

Finite VC Dim :- $d_{vc}(H_k)$

Derive & Prove :- tightest upper bound & lower bound on $d_{vc}\left(\bigcap_{k=1}^{K} H_k\right)$

↳ Now,

$$d_{vc}\left(\bigcap_{k=1}^{K} H_k\right) \geq 0 \qquad -(i)$$

↳ As it can't be negative integer. Also, it can be zero as the intersection can be an empty ϕ set.

↳ By intuition,

Let us assume that

$$d_{vc}\left(\bigcap_{k=1}^{K} H_k\right) > \min_k d_{vc}(H_k)$$

⇒ There are $a = d_{vc}\left(\bigcap_{k=1}^{K} H_k\right)$ # of points that intersection can shatter

⇒ Since, it's an intersection, the respective Hypothesis which was able to shatter must be present in all $H_k (k=1,\cdots K)$

⇒ $d_{vc}\left(\bigcap_{k=1}^{K} H_k\right) > \min_k d_{vc}(H_k)$ is a contradiction.

⇒ Hence, $d_{vc}\left(\bigcap_{k=1}^{K} H_k\right) \leq \min_k d_{vc}(H_k)$

Ans:-

$$0 \leq d_{vc}\left(\bigcap_{k=1}^{K} H_k\right) \leq \min_k d_{vc}(H_k)$$

5. AML Pb. 4.4 (a)-(c)

$x = [-1, 1]$ Uniform $p(x) = \frac{1}{2}$

$f(x) = \sum_{q=0}^{Q_f} a_q L_q(x)$, $L_q(x)$ ~ Legendre polynomials

$D = (x_1, y_1), \cdots (x_N, y_N)$, $y_n = f(x_n) + \sigma \varepsilon_n$

y_n & ε_n ~ i.i.d Normal R.V

Legendre polynomials:-

$$L_0(x) = 1$$

$$L_1(x) = x$$

$$L_k(x) = \frac{2k-1}{k} x L_{k-1}(x) - \frac{k-1}{k} L_{k-2}(x)$$

Specified Values:- Q_f, N, σ

a) We normalize f $\Rightarrow$ $E_{a,x}[f(x)] = 0$

&

$\text{Var}[f(x)] = E_{a,x}[f^2(x)] = 1$

$\Rightarrow E_{a,x}[y_n] = E_{a,x}[f(x_n) + \sigma \varepsilon_n]$

$= \underbrace{E_{a,x}[f(x_n)]}_{0} + \sigma \underbrace{E_{a,x}[\varepsilon_n]}_{0}$

$\Rightarrow E_{a,x}[y_n] = 0$

Also,

$$Var[y_n] = Var[f(x) + \sigma\varepsilon_n]$$

$$= Var[f(x)] + \sigma^2 Var[\varepsilon_n] + 2\sigma Cov(f(x), \varepsilon_n)$$

$$= 1 + \sigma^2 + 2\sigma Cov(f(x), \varepsilon_n)$$

If $f(x)$ & ε_n are not correlated.

$\Rightarrow Var[y_n] = 1 + \sigma^2 + 0$

$\Rightarrow$ For Different order of Polynomial, comparision beiomes easy bet" $f(x)$ & noise term.

b) $D = \{(x_1, y_1), (x_2, y_2) \cdots (x_N, y_N)\}$

$g_2 \rightarrow$ Best fit Hypothesis from H_2 [Second order polynomial]

$g_{10} \rightarrow$ Best fit Hypothesis from H_{10} [10^{th} order polynomial]

$\Rightarrow$ Comparing with Linear Regression model,

$$Y = \omega X \quad , \quad \omega^* = (X^T X)^{-1} X^T y$$

Here,

$$f(\underline{x}) = \sum_{q=0}^{Q_1} a_q L_q(x) = \phi_q(\underline{x})$$

$\Rightarrow$ Input Matrix $= \begin{bmatrix} \phi_Q(\underline{x}_1) \\ \phi_Q(\underline{x}_2) \\ \\ \phi_Q(\underline{x}_N) \end{bmatrix}_{(N)\times(Q+1)}$ ← Let it be represented by ϕ_Q

$\Rightarrow \quad w^* = (\phi_Q^T \phi_Q)^{-1} \phi_Q^T y$ ← By Least Squares Best fit

$\Rightarrow \quad g_Q(\underline{x}) = \phi_Q(\underline{x}) (\phi_Q^T \phi_Q)^{-1} \phi_Q^T y$

c) For a single experiment,

$$E_{out} = E_{x,y}\left[(g_{10}(\underline{x}) - y(\underline{x}))^2\right]$$

$$= E_{x,y}\left[g_{10}^2(\underline{x}) + y^2(\underline{x}) - 2\, g_{10}(\underline{x})\, y(\underline{x})\right]$$

$$= E_x\left[g_{10}^2(\underline{x})\right] + \underbrace{E_{x,y}\left[y^2(\underline{x})\right]}_{Var(y)} - 2\, E_{x,y}\left[g(\underline{x})\right] \underbrace{E_{x,y}\left[y(\underline{x})\right]}_{0}$$

$$= E_x\left[g_{10}^2(\underline{x})\right] + (1+\sigma^2) - 0$$

$$= E_x\left[g_{10}^2(\underline{x})\right] + (1+\sigma^2)$$

→ {g, f(x) & ε_n are not correlated}

5. (i) Fig 4.3(a), $\sigma^2 = 0.5$, $N = 60 \sim 130$.

↳ With fixed Noise level, $\sigma^2 = 0.5$, as we increase Data Points, $N = 60$ to 130, Overfitting Decreases.

↳ As the signal f is normalized, noise level is calibrated to signal & noise level σ^2 plays a major factor.

↳ With smaller N, & more complex, larger model, noise is more susceptible to deviate signal from true function.

↳ For a small Dataset, the model is less capable to distinguish noise from signal & hence tries to fit the noise, leading to overfit.

ii) Fig 4.3(a), $N = 100$, $\sigma^2 = 0$ to $\sigma^2 = 2$.

↳ With increase in Noise, Overfitting increases.

↳ As f is normalized i.e. $\text{Var}(f) = 1$ noise level is calibrated to signal & has major influence.

↳ More large, complex models (H_{10} compared to H_2) are more susceptible to noise as, it has more ways to deviate from true function. Hence, increase in overfitting.

(iii) Fig 4.3(b), $N = 75$, $Q_f = 0$ to $Q_f = 100$

↳ With higher target function complexity Q_f, it will be more susceptible to noise as it has more ways to deviate from true signal even though there is non-linearity.

↳ Hence with increase in Q_f (> 15), overfitting increases

↳ With $N = 75$, with lower range of target complexity, overfitting decreases ($Q_f = 0$ to $Q_f \approx 15$)

↳ With $N = 75$, & $Q_f > 15$, it overfits the data compared to H_2

↳ With $Q_f = 0$ to $Q_f = 15$, it seems like the signal was able to distinguish between noise & true function due to non-linearity of $Q_f(x)$. Hence overfitting decreases from $Q_f = 0$ to $Q_f = 15$.