ECE 595 HW 5

Sayeed Shafayet Chowdhury

Class ID - 19

1.(a)(i) We know the growth function for positive rays is $N+1$. If we enumerate the dichotomies added by negative rays, we get $N-1$ new dichotomies (opposite ones from pos. rays but we have to subtract the two dichotomies where al points +1 and all are -1). So, in total, $m_H(N) = 2N$

As the largest value of N for which $m_H(N) = 2^N$ is 2 ($\because m_H(3) = 6$), we have $\boxed{d_{vc} = 2}$

(ii) We know that the growth function for positive intervals is equal to $\frac{N^2}{2} + \frac{N}{2} + 1$.

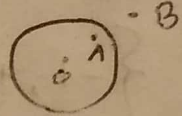
If we add the new dichotomies generated by negative intervals, we get $N-2$ new ones (for example for $N=3$, we only add the $(+1,-1,+1)$ dichotomy and for $N=4$, we add the $(+1,-1,+1,+1)$ and $(+1,+1,-1,+1)$ dichotomy)

Scanned with CamScanner

Of course, this only hold if $N > 1$, for $N = 1$ we already generate the two dichotomies with the positive intervals alone. In conclusion, we may write.

$$m_H(N) = \frac{N^2}{2} + \frac{3N}{2} - 1 \text{ if } N > \text{ and } 2 \text{ if } N = 1$$

As the largest value of N for which $m_H(N) = 2^N$ is $3 \left(\because m_H(4) = 13\right)$, we have $\boxed{d_{vc} = 3}$

(iii) Here, 

Say O is the center, if we have just one point A, we can either keep it inside the circle or outside. For $N = 2$, say point B is farther away from origin than A. Now the case $(A, B) = (-1, +1)$ is not possible since if we have to include B, A is auto-included. So, max^m no. of points that can be shattered is 1. So $\boxed{\text{VC. dim} = 1}$

(iv) It is same as (iii), just with a different center, this doesn't change the dichotomies, So $\boxed{d_{vc} = 1}$

(b) First we choose N points $x_1 = 10^1$, $x_2 = 10^2, \ldots$ $x_N = 10^N$ in $\mathbb{R}$, then we let $y = (y_1, \ldots, y_N)^T \in \{-1,+1\}^N$ be any dichotomy. Now we consider $\alpha = 0.d_1 d_2 \ldots d_N$ with the digit $d_i = 1$ if $y_i = -1$ and $d_i = 2$ if $y_i = +1$, then we have,

$$h_\alpha(x_k) = (-1)^{\lfloor \alpha \cdot 10^k \rfloor} = y_k$$

for all $k = 1, \ldots N$. We may now conclude that $\mathcal{H}(x_1, \ldots x_N) = \{-1,+1\}^N$ (or $m_H(N) = 2^N$) for all N and so $d_{vc}(H) = \infty$.

The samples are just 1D value $\in \mathbb{R}$, however with d_{vc} ∞, it implies model complexity is arbitrarily large. I think this is worse than perceptron, since in perceptron d_{vc} is $d+1$ for $\mathbb{R}^d$, so in this case it would be $= 2$, which allow simplification than a complex model (Occam's razor).

Ex. 2. (a) loss function $E_{aug}(h) = E_{in}(h) + \frac{\lambda}{N}\theta_h^T\theta_h$

$$= \frac{1}{N}\|A\theta_h - y\|_2^2 + \frac{\lambda}{N}\|\theta_h\|_2^2$$

setting $\nabla\theta E_{Aug} = 0 \Rightarrow \frac{2}{N}(A^TA\theta_D - A^Ty) + \frac{2\lambda\theta_D}{N} = 0$

$\Rightarrow \theta_D = (A^TA + \lambda I)^{-1}A^Ty$, now $y = A\theta_f + \epsilon$

$\therefore \theta_D = (A^TA + \lambda I)^{-1}A^T(A\theta_f + \epsilon)$

(b) $\theta_D = (A^TA + \lambda I)^{-1}A^T(A\theta_f + \epsilon)$

$= (A^TA + \lambda I)^{-1}A^TA\theta_f + (A^TA + \lambda I)^{-1}A^T\epsilon$

$= (A^TA + \lambda I)^{-1}(A^TA + \lambda I - \lambda I)\theta_f + (A^TA + \lambda I)^{-1}A^T\epsilon$

$= \theta_f - \lambda(A^TA + \lambda I)^{-1}\theta_f + (A^TA + \lambda I)^{-1}A^T\epsilon$

(c) (i) $\bar{g}(x) = E_D[g^D(x)]$

$= E_D[\theta_D^Tx] = E_D[x^T\theta_D]$

$= E_D[x^T(\theta_f - \lambda(A^TA + \lambda I)^{-1}\theta_f + (A^TA + \lambda I)^{-1}A^T\epsilon)]$

$= E_A[x^T\theta_f - \lambda x^T(A^TA + \lambda I)^{-1}\theta_f + x^T(A^TA + \lambda I)^{-1}A^T\underbrace{E[\epsilon|A]}_{=0}]$

$\Rightarrow \bar{g}(x) = x^T\theta_f - \lambda x^T E_A[(A^TA + \lambda I)^{-1}]\theta_f$

$$= \theta_f^T x - \lambda x^T E_D\left[(A^TA+\lambda I)^{-1}\right]\theta_f$$

(ii) $\left(\bar{g}(x) - f(x)\right)^2$, here $f(x) = \theta_f^T x$

$$= \lambda^2 \theta_f^T E_A\left[(A^TA+\lambda I)^{-1}\right] x x^T E_A\left[(A^TA+\lambda I)^{-1}\right]\theta_f$$

$$= \lambda^2 \operatorname{trace}\left(x x^T E_A\left[(A^TA+\lambda I)^{-1}\right]\theta_f \theta_f^T E_A\left[(A^TA+\lambda I)^{-1}\right]\right)$$

[using cyclic property of trace]

(iii) bias $= E_x[\text{bias}(x)] =$

$$= \lambda^2 \operatorname{trace}\left(\underbrace{E_x\left[x x^T\right]}_{=I} E_A\left[(A^TA+\lambda I)^{-1}\right]\theta_f \theta_f^T E_A\left[(A^TA+\lambda I)^{-1}\right]\right)$$

$$= \lambda^2 \operatorname{trace}\left(E_A\left[(A^TA+\lambda I)^{-1}\right]\theta_f \theta_f^T E_A\left[(A^TA+\lambda I)^{-1}\right]\right)$$

Now using $A^TA \approx N E_x[x x^T] = N I$

So, $E_A\left[(A^TA+\lambda I)^{-1}\right] \simeq E_A\left[\left(\frac{1}{N+\lambda} I\right)\right] = \frac{1}{N+\lambda} I$

So, bias $\approx \frac{\lambda^2}{(N+\lambda)^2} \underbrace{\operatorname{trace}\left(\theta_f \theta_f^T\right)}_{= \operatorname{trace}(\theta_f^T \theta_f) = \|\theta_f\|^2}$

$$\approx \frac{\lambda^2}{(N+\lambda)^2} \|\theta_f\|_2^2$$

(iv) $var(x) = E_D\left[(h^D - \bar{g}(x))^2\right]$

$$= E_D\left[\lambda x^T\left(E_A\left[(A^TA+\lambda I)^{-1}\right] - (A^TA+\lambda I)^{-1}\right)\theta_f + x^T(A^TA+\lambda I)^{-1}A^T\epsilon)^2\right]$$

Now $E_A\left[(A^TA+\lambda I)^{-1}\right] \approx \frac{1}{N+\lambda} I$, also $(A^TA+\lambda I)^{-1} = \frac{1}{N+\lambda} I$

So, $var(x) \approx E_D\left[\epsilon^T A(A^TA+\lambda I)^{-1} x x^T (A^TA+\lambda I)^{-1} A^T \epsilon\right]$

$$\approx E_A\left[trace(\underbrace{E_{y|A}[\epsilon\epsilon^T]}_{=\sigma^2 I} A(A^TA+\lambda I)^{-1} x x^T (A^TA+\lambda I)^{-1} A^T\right]$$

$$\approx \sigma^2 E_A\left[trace\left(x x^T (A^TA+\lambda I)^{-1} A^TA (A^TA+\lambda I)^{-1}\right)\right]$$

[using cyclic property of trace]

(v) $var = E_x\left[var(x)\right]$

$$\approx \sigma^2 E_A\left[trace\left(\underbrace{E_x[x x^T]}_{=I}(A^TA+\lambda I)^{-1} A^TA (A^TA+\lambda I)^{-1}\right)\right]$$

$$\approx \sigma^2 E_A\left[trace\left(\underbrace{I}_{I \approx \frac{1}{N}A^TA}(A^TA+\lambda I)^{-1} A^TA (A^TA+\lambda I)^{-1}\right)\right] \quad [\because A^TA \approx NI]$$

$$\approx \frac{\sigma^2}{N} E_A\left[trace\left(A(A^TA+\lambda I)^{-1} A^T \underbrace{A(A^TA+\lambda I)^{-1} A^T}_{H(\lambda)}\right)\right]$$

$$\approx \frac{\sigma^2}{N} E_A\left[trace(H^T H)\right] = \frac{\sigma^2}{N} E_A\left[trace(H(\lambda)^2)\right]$$

(d) Asymptotic properties -

when $\lambda = 0 \Rightarrow$ bias $\approx \frac{0}{N^2} \|\theta_f\|^2 \approx 0$

$\text{var} = \frac{\sigma^2}{N} E_A[\text{tr}(I)] = \frac{\sigma^2}{N}(d+1)$

Again, as $\lambda \to \infty$, $N + \lambda \to \lambda$

bias $\approx \frac{\lambda^2}{\lambda^2} \|\theta_f\|^2 \approx \|\theta_f\|_2^2$

var $\approx \frac{\sigma^2}{N} E_A[0] \approx 0$

with low regularization ($\lambda \to 0$), the model overfits (as it fits training data nearly perfectly), so bias ≈ 0, but it is unlikely to fit well to new data (so verey sensitive to deviations in training set) $\to$ so high variance. With lareger & lareger λ, bias increases but variance ≈ 0, so reg. tries to reduce the var of estimator by simplifying it. Again for N, no. of

$N \to 0$, bias $\approx \|\theta_f\|_2^2$, var $\approx \infty$ with very low training data we have both high bias & high var. with $N \to \infty$ bias ≈ 0, var ≈ 0, so with infinite no. of sample we can both fit model perfectly & generalise well.