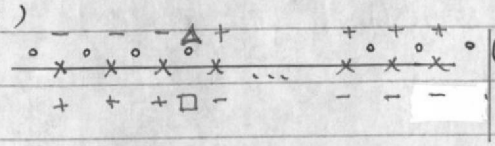
1.

Problem 2.3 Compute the maximum number of dichotomies, $m_H(N)$, for these learning models, and consequently compute d_{vc}, the VC dimension.

(a) Positive or negative ray: H contains the functions which are +1 on $[a, \infty)$ (for some a) together with those that are +1 on $(-\infty, a]$ (for some a)



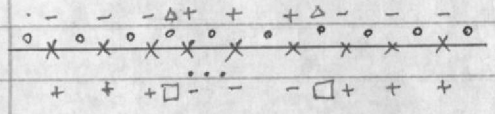
$(N+1)$ (o) selecting points
N (x) data points
△ positive ray
□ negative ray

$m_H(N) = N+1$

break point $k = 2 = d_{vc} + 1$

$d_{vc} = 1$

(b) Positive or negative interval: H contains the functions which are +1 on an interval $[a,b]$ and -1 elsewhere or -1 on an interval $[a,b]$ and +1 elsewhere.



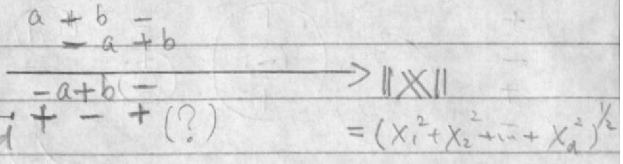
$(N+1)$ (o) selecting points
N (x) data points
△ positive interval
□ negative interval

$$m_H(N) = \binom{N+1}{2} + 1$$
$$= \frac{(N+1)(N)(N-1)!}{2!\ (N-1)!} + 1$$
$$= \frac{1}{2}(N^2+N) + 1$$
$$= \frac{N^2}{2} + \frac{N}{2} + 1$$

break point $k = 3 = d_{vc} + 1$

$d_{vc} = 2$

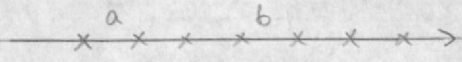
(c) Two concentric spheres in $\mathbb{R}^d$: H contains the functions which are +1 for $a \le (x_1^2 + \cdots + x_d^2)^{1/2} \le b$



in this case, no data sets can be shattered by the spheres in $\mathbb{R}^d$

⇒ break point $k = 3 = d_{vc} + 1$

$d_{vc} = 2$



$$m_H(N) = \binom{N+1}{2} + 1$$
$$= \frac{N^2}{2} + \frac{N}{2} + 1$$

$$1 + N + \frac{N^3 - 3N^2 + 2N}{6}$$

2. Problem 2.8 Which of the following are possible growth function $m_H(N)$ for some hypothesis set: $1+N$; $1+N+\frac{N(N-1)}{2}$; 2^N; $2^{\lfloor \sqrt{N} \rfloor}$; $2^{\lfloor N/2 \rfloor}$; $1+N+\frac{N(N-1)(N-2)}{6}$

$1+N$ is possible, e.g. $m_H(N)$ of positive ray.

$1+N+\frac{N(N-1)}{2}$ is possible, e.g. $m_H(N)$ of positive interval

2^N is possible, e.g. $m_H(N)$ of convex set.

$2^{\lfloor \sqrt{N} \rfloor}$ and $2^{\lfloor N/2 \rfloor}$ are not possible because they're not monotonically increasing function.

3. Problem 2.10 Show that $m_H(2N) \leq m_H(N)^2$, and hence obtain a generalization bound which only involves $m_H(N)$.

If we seperate $2N$ to N and N,

each $m_H(N) \leq 2^N$

and if these two set are independent

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

if in other dependent cases

$m_H(2N) < m_H(N)^2$ because there're some of possible dicotomies in $m_H(2N)$ interseeit with those in $m_H(N)^2$.

Therefore $m_H(2N) \leq m_H(N)^2$

4. Problem 2.13

(a) Let $\mathcal{H}=\{h_1, h_2, \ldots, h_M\}$ for a finite M. Prove that $d_{vc}(\mathcal{H}) \le \log_2 M$

(b) For hypothesis sets $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_K$ with finite VC dimension $d_{vc}(\mathcal{H}_k)$, derive and prove the tightest upper and lower bound that you can get on $d_{vc}(\bigcap_{k=1}^{K} \mathcal{H}_k)$

(c) For hypothesis sets $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_K$ with finite VC dimension $d_{vc}(\mathcal{H}_k)$, derive and prove the tightest upper and lower bounds that you can get on $d_{vc}(\bigcup_{k=1}^{K} \mathcal{H}_k)$

(a) $M \le m_{\mathcal{H}}(N) \le 2^k$ $d_{vc} \to \infty$

k : break point

$m_{\mathcal{H}}(N) \le \sum_{i=0}^{d_{vc}} \binom{N}{i}$

$m_{\mathcal{H}}(N) \le (N)^{d_{vc}} + 1$

$k-1$ points shattered

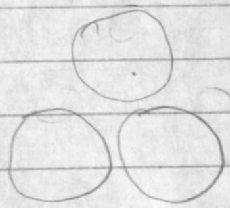
$B(N_0+1, k) \le B(N_0, k) + B(N_0, k-1)$

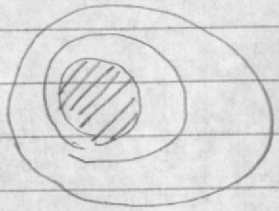
$2^{k-1} \le m_{\mathcal{H}}(N) \le 2^k$

$2^{d_{vc}} \le m_{\mathcal{H}}(N) \le M$

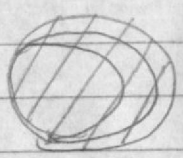
$\Rightarrow d_{vc}(\mathcal{H}) \le \log_2 M$.

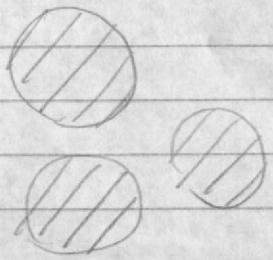
(b) $\phi \le d_{vc}(\bigcap_{k=1}^{K} \mathcal{H}_k) \le \min\{d_{vc}(\mathcal{H}_k)\}_{k=1}^{K}$





(c) $\max\{d_{vc}(\mathcal{H}_k)\}_{k=1}^{K} \le d_{vc}(\bigcup_{k=1}^{K} \mathcal{H}_k) \le \sum_{k=1}^{K} d_{vc}(\mathcal{H}_k)$





5. Problem 2.22 When there is noise in data, $E_{out}(g^{(D)}) = \mathbb{E}_{x,y}[(g^{(D)}(x) - y(x))^2]$, where $y(x) = f(x) + \epsilon$, If ϵ is a zero-mean noise random variable with variance σ^2, show that the bias-variance decomposition becomes $\mathbb{E}_D[E_{out}(g^{(D)})] = \sigma^2 + \text{bias} + \text{var}$

$$\mathbb{E}_D[E_{out}(g^{(D)})] = \mathbb{E}_D[\mathbb{E}_{x,y}[(g^{(D)}(x) - y(x))^2]]$$
$$= \mathbb{E}_D[\mathbb{E}_{x,y}[(g^{(D)}(x) - (f(x)+\epsilon))^2]]$$
$$= \mathbb{E}_{x,y}[\mathbb{E}_D[g^{(D)}(x)^2 - 2g^{(D)}(x)f(x) - 2g^{(D)}(x)\epsilon + f(x)^2 + 2f(x)\epsilon + \epsilon^2]]$$

$$= \mathbb{E}_{x,y}[\mathbb{E}_D[g^{(D)}(x)^2] - 2\mathbb{E}_D[g^{(D)}(x)]f(x) + f(x)^2$$
$$- 2\mathbb{E}_D[g^{(D)}(x)]\mathbb{E}_D[\epsilon] + 2f(x)\mathbb{E}_D[\epsilon] + \mathbb{E}_D[\epsilon^2]$$
$$= \mathbb{E}_{x,y}[\mathbb{E}_D[g^{(D)}(x)^2] - \bar{g}(x)^2$$
$$+ \bar{g}(x)^2 - 2\bar{g}(x)f(x) + f(x)^2$$
$$+ 0 + 0 + \mathbb{E}_D[\epsilon^2]]$$
$$= \text{Var} + \text{bias} + \sigma^2$$

6. $X \in \mathbb{R}^1 : [-1, 1]$, data set: $\{x_1, x_2\}$, target function $f(x) = x^2$

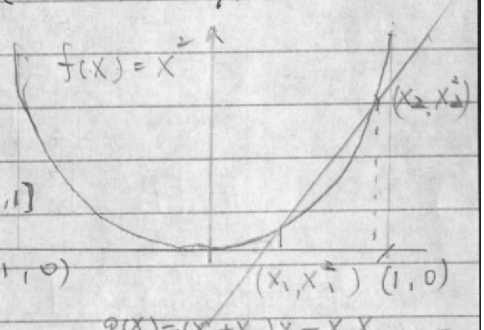
Problem 2.24 $D = \{(x_1, x_1^2), (x_2, x_2^2)\}$. The learning Algorith returns the line fitting these two points as g ($\mathcal{H}$ consists of functions of the form $h(x) = ax + b$) We are interested in the best performance ($\mathbb{E}[E_{out}]$) of our learning system with respect to the squared error measure, the bias and the var

(a) $\bar{g}(x)$ analytic expression

(b) Experiment that you could run to determine (numerically) $\bar{g}(x)$, $\mathbb{E}[E_{out}]$, bias, and var.

(c) Run your experiment and report the results. Compare $\mathbb{E}[E_{out}]$ (expectation is with respect to data set) with bias+var. Provide a plot of your $\bar{g}(x)$ and $f(x)$ (on the same plot)

(d) Compute analytically what $\mathbb{E}[E_{out}]$, bias and var should be



(a) $\bar{g}(x) = \bar{a}x + \bar{b}$, where $\bar{a} = \mathbb{E}[x_1 + x_2]$

$= \mathbb{E}[x_1] + \mathbb{E}[x_2] = 0 + 0 = 0$, $\because x_1, x_2 \in [-1, 1]$

$\bar{b} = \mathbb{E}[x_1 x_2] = \mathbb{E}[x_1]\mathbb{E}[x_2] = 0 \cdot 0 = 0$

$\because x_1$ & x_2 are independent $\Rightarrow \bar{g}(x) = 0$

(b) The experiment would be:

#1 step: Generating K (input variable, e.g. 100) numbers of data set $D_1, \ldots, D_K$, where $D = \{(x_1, x_1^2), (x_2, x_2^2)\}$ and x_1, x_2 are generated randomly $[-1, 1)$, in implementation D would be a $(K \text{ by } 4)$ matrix.

#2 step: Compute $\bar{g}(x) = \bar{a}x + \bar{b}$, where $\begin{cases} \bar{a} = \frac{1}{K}\sum_{i=1}^{K} (x_1^{(D_i)} + x_2^{(D_i)}) = \\ \bar{b} = \frac{1}{K}\sum_{i=1}^{K} (x_1^{(D_i)} x_2^{(D_i)}) \end{cases}$ and function $g(x) = ax + b$

#3 step: Compute bias $= \frac{1}{2}\int_{-1}^{1} ((\bar{a}x + \bar{b}) - x^2)^2 dx$

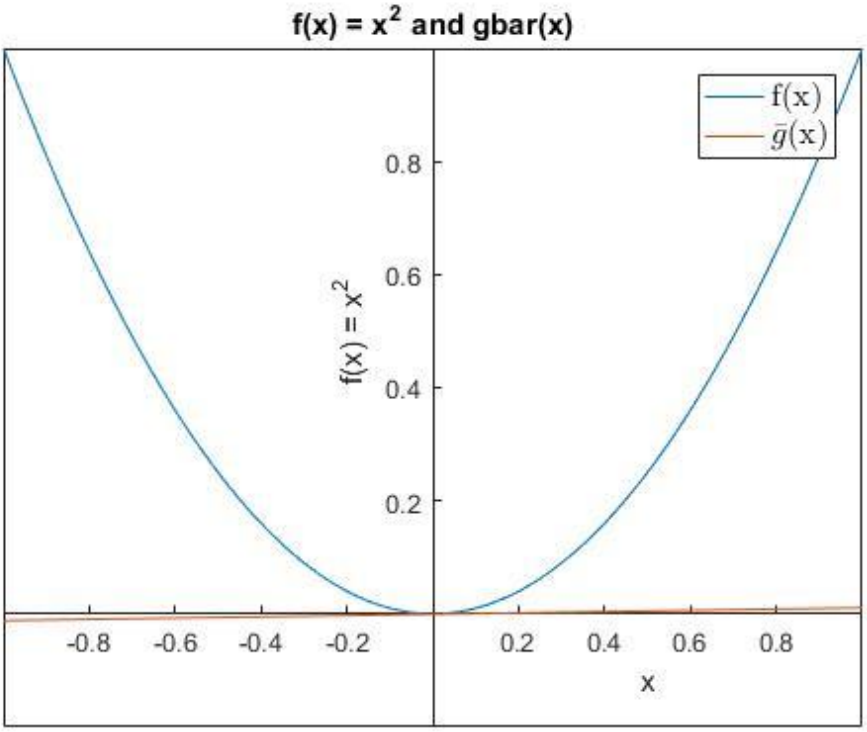
Compute Var $= \mathbb{E}_x\left[\frac{1}{K}\sum_{i=1}^{K} (g(x_1^{D_i}) - \bar{g}(x_1^{D_i}))^2 ; \frac{1}{K}\sum_{i=1}^{K} (g(x_2^{D_i}) - \bar{g}(x_2^{D_i}))^2\right]$

$\mathbb{E}_x[\text{mean}[(x_1^{D_i} + x_2^{D_i})x_1^{D_i} - (x_1^{D_i} x_2^{D_i}) - (\bar{a}x_1^{D_i} + \bar{b}))^2]$;

$\text{mean}[(x_1^{D_i} + x_2^{D_i})x_2^{D_i} - (x_1^{D_i} x_2^{D_i}) - (\bar{a}x_2^{D_i} + \bar{b}))^2]]$

$$E_D[E_{out}] = \frac{1}{5} + \frac{1}{3}E_D[(x_1 + x_2)^2] - \frac{2}{3}E_D[x_1 x_2] + E_D[(x_1 x_2)^2]$$

Problem 2.24 (c)



	Analytical computation (from part(d))	Numerical computation (experiment with k=10000)
$\mathrm{E}[E_{out}]$	$\frac{8}{15} \cong 0.5333$	0.5274
Bias + Var	$\frac{8}{15} \cong 0.5333$	0.53012
Bias	$\frac{1}{5} = 0.2$	0.2004
Var	$\frac{1}{3} \cong 0.3333$	0.3297
$\bar{g}(x) = \ \bar{a}x + \ \bar{b}$	0	$0.0113x - 0.0006$

Another way to compute Var is on the next page.

(d) $\mathrm{Var}(X_1) = \mathbb{E}_D[X_1^2] - (\mathbb{E}_D(X_1))^2 = \frac{1}{3} - 0 = \frac{1}{3}$

$\mathrm{Var}(X_2) = \mathbb{E}_D[X_2^2] - (\mathbb{E}_D(X_2))^2 = \frac{1}{3} - 0 = \frac{1}{3}$

$\mathbb{E}_D[X^2] = \int_{-1}^{1} \frac{1}{2} x^2 dx = \frac{1}{3}$

$\mathbb{E}_D[X] = 0$

$\mathrm{Var} = \frac{1}{2}[\mathrm{Var}(X_1) + \mathrm{Var}(X_2)] = \frac{1}{3}$

$E_{out} = \int_{-1}^{1} \frac{1}{2}[g(x) - f(x)]^2 dx$

$= \int_{-1}^{1} \frac{1}{2}[(ax+b) - x^2]^2 dx$

$= \int_{-1}^{1} \frac{1}{2}[x^4 - 2ax^3 + (a^2 - 2b)x^2 + 2abx + b^2] dx$

$= \frac{1}{5} + \frac{a^2 - 2b}{3} + b^2$

$\mathbb{E}_D[E_{out}] = \frac{1}{5} + \frac{1}{3}\mathbb{E}_D[(X_1 + X_2)^2] - \frac{2}{3}\mathbb{E}_D[X_1 X_2] + \mathbb{E}_D[(X_1 X_2)^2]$

$\mathbb{E}_D[X_1 X_2] = \mathbb{E}_D[X_1]\mathbb{E}_D[X_2] = 0 \cdot 0 = 0$

$\mathbb{E}_D[X_1 + X_2] = \mathbb{E}_D[X_1] + \mathbb{E}_D[X_2] = 0 + 0 = 0$

$\mathbb{E}_D[(X_1 + X_2)^2] = \mathbb{E}_D[X_1^2] + 2\mathbb{E}_D[X_1]\mathbb{E}_D[X_2] + \mathbb{E}_D[X_2^2]$

$= \frac{1}{3} + 2 \cdot 0 \cdot 0 + \frac{1}{3}$

$= \frac{2}{3}$

$\Rightarrow \mathbb{E}_D[E_{out}] = \frac{1}{5} + \frac{1}{3} \cdot \frac{2}{3} - \frac{2}{3}(0) + \frac{1}{3} \cdot \frac{1}{3}$

$= \frac{8}{15}$

$\mathrm{Bias} = \int_{-1}^{1} \frac{1}{2}[\bar{g}(x) - f(x)]^2 dx$

$= \frac{1}{5} + \frac{\bar{a}^2 - 2\bar{b}}{3} + \bar{b}^2$, from result of (a), $\bar{a} = \bar{b} = 0$

$= \frac{1}{5}$

6. Problem 2.24(d)

$$Var = E_X[E_b[(g^b(x) - \bar{g}(x))^2]]$$

$$E_X[E_a[E_b[(ax+b)^2]]]$$

$$E_X[E_b[a^2X^2 + 2abx + b^2]]$$

$$E_X[X^2 E_b[(X_1+X_2)^2] + X \cdot 2E_b[X_1+X_2]E_a[X_1X_2] + E_a[X_1^2X_2^2]]$$

$$E_X[\tfrac{2}{3}X^2 + 2X \cdot 0 \cdot 0 + \tfrac{1}{9}]$$

$$= \int_{-1}^{1} \tfrac{1}{2}\left[\tfrac{2}{3}x^2 + \tfrac{1}{9}\right]dx$$

$$= \tfrac{1}{2}\left[\tfrac{2}{9}x^3 + \tfrac{1}{9}x\right]\Big|_{-1}^{1}$$

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

$$= \tfrac{1}{3}$$

7. $H = X(X^TX)^{-1}X^T$, X ~~$\in \mathbb{R}$~~ is an N by $d+1$ matrix, X^TX invertible

Exercise 3.3(a) Show that H is symmetric

first, since X^TX is invertible $\Rightarrow [(X^TX)^{-1}]^T = [(X^TX)^T]^{-1}$

$= (X^TX)^{-1}$ — (1)

Second, $H^T = [X(X^TX)^{-1}X^T]^T$

$= (X^TX)^{-1}X^T)^T X^T$

$= X[(X^TX)^{-1}]^T X^T$, using result (1)

$= X(X^TX)^{-1}X^T$

$= H$

$\Rightarrow H$ is symmetric

(b) Show that $H^k = H$ for any positive integer K

$k=1$, $H^1 = H$

$k=2$, $H^2 = H \cdot H = [X(X^TX)^{-1}X^T][X(X^TX)^{-1}X^T]$

$= X(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^T$

$= X(X^TX)^{-1}X^T$

$= H$

$n \in \mathbb{Z}^+$, $k=n$, $H^n = \underbrace{H \cdot H \cdot \ldots \cdot H}_{n} = \underbrace{(X(X^TX)^{-1}X^T)(X(X^TX)^{-1}X^T)\ldots(X(X^TX)^{-1}X^T)}_{n}$

$= X \cdot \underbrace{I \cdot I \cdot \ldots \cdot I}_{n-1} \cdot (X^TX)^{-1}X^T$, where I is $(d+1)$ by $(d+1)$ identity matrix

$= X(X^TX)^{-1}X^T$

Therefore, $H^k = H$, $k \in \mathbb{Z}^+$

(c) If I is the identity matrix of size N, show that $(I-H)^k = I-H$ for any positive integer K.

$k=1$, $(I-H)^1 = (I-H)$

$k=2$, $(I-H)^2 = (I-H)(I-H) = I^2 - IH - HI + H^2$ (using result from (b))

$= I - H - H + H = (I-H)$

$k=n$, $(I-H)^n = \underbrace{(I-H)(I-H)\cdots(I-H)}_{n}$

$= \underbrace{(I-H)(I-H)\cdots(I-H)}_{n-1}$

$\vdots$

$= (I-H)$