# hw 5

yasir Mohamed

October 2024

# 1 Exercise 2.8

## 1.1 A

We can assume H contain the set of hypothesis H = $g_1, g_2 ... g_n$ then we can define our $\overline{g} = \frac{1}{N}\sum_{i=1}^{N}\frac{g_i}{N}$ which is a linear combination of the hypothesis in the hypothesis set and by definition it should be included in H

## 1.2 B

H = [+1, -1]. Where the first hypothesis always returns +1 and the second always returns -1. The average hypotheses set

$$\overline{g} = \frac{1}{2}\sum_{n=1}^{2} g_n(x)$$

does not exist in the hypothesis set

## 1.3 C

$\overline{g}$ can be a binary function but that is most likely not the case. if the hypothesis set only contains the hypothesis that returns +1 or -1 then it returns a binary hypothesis. otherwise, it will return a nonbinary hypothesis

## 2 Problem 2.14

### 2.1 A

Given that all $H_n \in H$ have the same finite VC dimension $d_{VC}$ then each is able to form exactly $d_{VC}$ unique dichotomies without shatter.

Now given that H is the combined set of all the maximum number of dichotomies it can contain is achieved if each dichotomy in every $H_n$ is unique. Resulting in $k * d_(vc)$ unique dichotomies without shattering therefore

$$d_{VC}(H) \leq Kd_{VC} < k(d_{VC}+1)$$

### 2.2 B

Needs to prove the following:

$$2^l > 2kl^{d_{vc}} > m(l) \leq l^{(d_{VC})} + 1$$

$$2kl^{d_{vc}} \overset{?}{>} l^{d_{VC}} + 1 < k(l^{d_{VC}} + 1) = kl^{d_{VC}} + k$$

$$2kl^{d_{vc}} \overset{?}{>} kl^{d_{VC}} + k$$

$$2kl^{d_{vc}} = kl^{d_{vc}} + kl^{d_{vc}} \overset{?}{>} kl^{d_{VC}} + k$$

given that $d_{VC} \geq 0$ then $kl^{d_{vc}} > k$ :

$$2kl^{d_{vc}} = kl^{d_{vc}} + kl^{d_{vc}} > kl^{d_{VC}} + k > m(l)$$

$$2kl^{d_{VC}} > m(l)$$

and as such

$$d_{VC}(H) \leq l$$

### 2.3 C

$$d_{VC} \leq min(k(d_{VC}+1), 7(d_{VC}+K)log_2(d_{VC}k))$$

case one $min(k(d_{VC}+1), 7(d_{VC}+K)log_2(d_{VC}k)) = k(d_{VC}+1)$:

$$d_{VC} \leq k(d_{VC}+1)$$

this was proven in part A

case two $min(k(d_{VC}+1), 7(d_{VC}+K)log_2(d_{VC}k)) = 7(d_{VC}+K)log_2(d_{VC}k))$: Building up from section B if we need to prove that with l = $7(d_{VC}+K)log_2(d_{VC}k))$ the inequality $2^l > 2kl^{d_{VC}}$ still holds as that proves that $l \leq d_{VC}$

$$2^{7(d_{VC}+K)log_2(d_{VC}k))} > 2k(7(d_{VC}+K)log_2(d_{VC}k)))^{d_{VC}}$$

Take log base 2:

$$7(d_{VC}+K)log_2(d_{VC}k)) > log_2(2k(7(d_{VC}+K)log_2(d_{VC}k)))^{d_{VC}}$$

$$= log_2(2) + log_2(k) + d_{VC}log_2(7(d_{VC}+K)log_2(d_{VC}k))))$$

$$7(d_{VC}+K)log_2(d_{VC}k)) > log_2(2k) + d_{VC}log_2(7(d_{VC}+K)log_2(d_{VC}k))$$

we know that both k and $d_{VC}$ are $> 1$:

$$log_2(2k)+d_{VC}log_2(7(d_{VC}+K)log_2(d_{VC}k)) < log_2(d_{VC}k)+d_{VC}log_2(7(d_{VC}+K)log_2(d_{VC}k))$$

$$< log_2(d_{VC}k) + d_{VC}log_2(7(d_{VC}+K)) + d_{VC}log_2(log_2(d_{VC}k))$$

$$< log_2(d_{VC}k) + d_{VC}log_2(7(d_{VC}+K)) + d_{VC}log_2(d_{VC}k)$$

$$= log_2(d_{VC}k) + d_{VC}log_2(7) + d_{VC}log_2(d_{VC}+K)) + d_{VC}log_2(d_{VC}k)$$

$$= log_2(d_{VC}k) + d_{VC}log_2(7) + 2d_{VC}log_2(d_{VC}+K))$$

we know that $log_2(d_{VC}k) < 7klog_2(d_{VC}k)$ since k $> 1$

$$< 7klog_2(d_{VC}k) + d_{VC}log_2(7) + 2d_{VC}log_2(d_{VC}+K))$$

We also know that the $log_2(7) < 5log_2(d_{VC}K)$ since both $d_{VC}$ and k are bigger than one resulting in a larger log.

$$< 7klog_2(d_{VC}k) + 5d_{VC}log_2(d_{VC}k) + 2d_{VC}log_2(d_{VC}+K))$$

$$= 7klog_2(d_{VC}k) + 7d_{VC}log_2(d_{VC}k) = 7(d_{VC}+K)log_2(d_{VC}k))$$

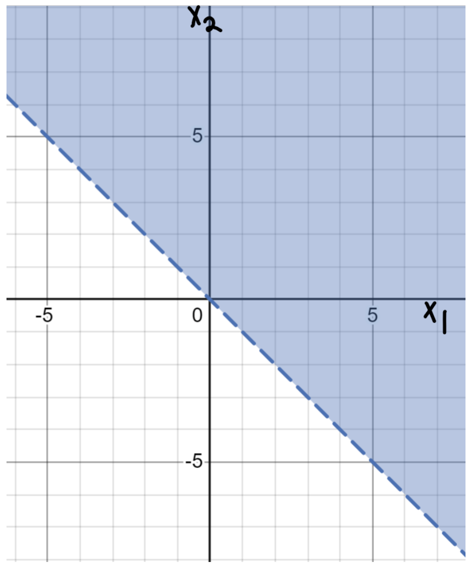


Figure 1: h(x) with the +1 region in blue

# 3 Problem 2.15

## 3.1 A

$$h(x) = sign(x_1 + x_2)$$

Here if you increase either $x_1$ or $x_2$ , the sum $x_1 + x_2$ will either stay the same or increase, ensuring that the classifier respects the ordering of inputs.

This can be represented using the line $x_1 = -x_2$ shown below with the blue region representing $+1$ and the white region representing -1

## 3.2 B

the first point $x^1$ may have any values for the components$x_1^1$ and $x_2^1$ and can be classified as either $+1$ or -1.

From there on we can generate N points using the following format ($x_1^n = x_1^{n-1} - 1$, $x_2^n = x_2^{n-1} - 1$) as per the hint. Given that for no two points does the inequality $x_i \geq x > j$ hold for all their components( since we are increasing the first component and decrementing the second) the additional requirement on H($h(x_1) \geq h(x_2)$) does not apply. Therefore, any point n may be represented as either $+1$ or -1 regardless of the previous points.

# 4 Problem2.24

## 4.1 A

g(x) must follow the data set given any x value and as such we can confirm that:

$$g(x) = ax + b$$

eq1

$$g(x_1) = ax_1 + b = x_1^2$$

eq2

$$g(x_2) = ax_2 + b = x_2^2$$

from eq1 solve for a:

$$ax_1 = x_1^2 - b$$

$$a = \frac{x_1^2}{x_1} - \frac{b}{x_1} = x_1 - \frac{b}{x_1}$$

from eq2 solve for a:

$$ax_2 = x_2^2 - b$$

$$a = \frac{x_2^2}{x_2} - \frac{b}{x_2} = x_2 - \frac{b}{x_2}$$

we can then set them equal to each other and solve for b

$$x_1 - \frac{b}{x_1} = x_2 - \frac{b}{x_2}$$

$$x_1 - x_2 = -\frac{b}{x_2} + \frac{b}{x_1}$$

$$x_1 - x_2 = b(\frac{1}{x_1} - \frac{1}{x_2})$$

$$b = \frac{x_1 - x_2}{\frac{1}{x_1} - \frac{1}{x_2}} = \frac{x_1 - x_2}{\frac{x_2 - x_1}{x_1 x_2}} = \frac{(x_1 x_2)(x_1 - x_2)}{x_2 - x_1}$$

$$\frac{x_1 - x_2}{x_2 - x_1} = -1$$

$$b = -x_1 x_2$$

Next, we do the same but solve for b first, eq1:

$$b = x_1^2 - ax_1$$

eq2:

$$b = x_2^2 - ax_2$$

set them equal and solve for a:

$$x_1^2 - ax_1 = x_2^2 - ax_2$$

$$x_1^2 - x_2^2 = ax_1 - ax_2$$
$$x_1^2 - x_2^2 = a(x_1 - x_2)$$
$$a = \frac{x_1^2 - x_2^2}{x_1 - x_2} = x_1 + x_2$$

the final hypnosis is

$$g^D(x_n) = ax_n + b = (x_1 + x_2)x_n - x_1 x_2$$

Next to find the average:

$$\overline{g}(x) = E_D[g^D(x)]$$
$$\overline{g}(x) = E_D[ax + b]$$
$$\overline{g}(x) = xE_D[a] + E_D[b]$$

plug the a and b values:

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

$$\overline{g}(x) = xE_D[x_1] + xE_D[x_2] - E_D[x_1^2]E_D[x_2^2]$$

x is uniformly distributed over [-1,1], E[x] = 0

$$\overline{g}(x) = x(0) + x(0) - (0)(0) = 0$$

## 4.2 B

We can repeatedly take two random samples ranging from [-1,1] N times and then use those samples to calculate g(x) = ax+b -> $g^D(x) = ax + b$. we can then sum the values of g(x) and divide by N to get the $\overline{g}(x)$

$$\overline{g}(x) = \frac{1}{N}\sum_{i=1}^{N} g_i(x)$$

For variance we calculate

$$\frac{1}{N}\sum_{i=0}^{N}[g^D(x) - \overline{g}(x)]^2$$

for each g(x) calculated earlier. This will be done by averaging out the output of $\frac{1}{N}\sum_{i=0}^{N}[g^D(x) - \overline{g}(x)]^2$ over 1000 randomly selected samples for x

For bias, we could calculate

$$E[(\overline{g}(x) - f(x))^2]$$

by using 1000 samples to test the mean function compared to $f(x) = x^2$ and averaging that out by dividing by N

For the estimate of out-of-sample error, we can simply calculate the out-of-sample error for each g(x) generated over 1000 samples and average them out to get the total out-of-sample error

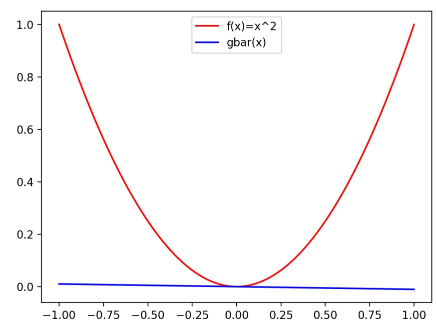


Figure 2: f(x) vs $\overline{g}(x)$

### 4.3 C

out of sample error: 0.5459006795594628

bias: 0.20050422830192013

variance: 0.3384489632671064

variance + bias: 0.5389531915690265

The variance + bias is very close to the value of out-of-sample error.

### 4.4 D

bias:

$$(\overline{g}(x) - f(x))^2 = \frac{1}{2}\int_{-1}^{1}(x^2)^2 dx = \frac{x^5}{5*2}\,|_{-1}^{1} = \frac{1}{10} - \frac{1}{10} = \frac{2}{10} = \frac{1}{5}$$

var:

$$E_x[E_D[(g^D(x) - \overline{g}(X))^2]] = \frac{1}{2}\int_{-1}^{1}\frac{1}{2}\int_{-1}^{1}\frac{1}{2}\int_{-1}^{1}((x_1+x_2)x - x_1x_2)^2 dx_1 dx_2 dx$$

$$= \frac{1}{8}\int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1}((x_1+x_2)x - x_1x_2)^2 dx_1 dx_2 dx$$

$$= \frac{1}{8}\int_{-1}^{1}\int_{-1}^{1}(\frac{2}{3} + x_2^2)x^2 - \frac{4}{3}x_2 + \frac{2}{3}x_2^2 dx_2 dx$$

$$= \frac{1}{8}\int_{-1}^{1}\frac{4}{3}x^2\frac{4}{9}dx$$

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

out of sample error = bias + variance

$$\frac{1}{3} + \frac{1}{5} = \frac{8}{15}$$