# Assignment 5

Name: Mohaiminul Al Nahian
RIN: 662026703
Course no.: CSCI 6100

## 1 Exercise 2.8

(a) If there are $D_1, D_2, ...., D_k$ datasets and the learning algorithm is applied to produce hypothesis $g_1, g_2, ..., g_k$ then the average function for estimating x is $\bar{g} \approx \frac{1}{K}\sum_1^k g_k(x) = \frac{1}{K}g_1 + \frac{1}{K}g_2 + ... + \frac{1}{K}g_k$. Since every $g_i$ is under H and it is given that H is closed under linear combination, so $\frac{1}{K}\sum_1^k g_k(x)$ is also in H, or $\bar{g}\epsilon H$.

(b) If H contains only two hypothesis $h_1 = g_1 = +1$ and $h_2 = g_2 = -1$, both with equal probability 0.5, then expected value, $\bar{g} = (+1) \times 0.5 + (-1) \times 0.5 = 0$, which is not in H.

(c) The binary classification cannot guarantee $\bar{g}$ to be binary. If H contains only one binary hypothesis $h_1$, then $\bar{g}$ is also binary. But let's say H contains multiple binary hypotheses, then $\bar{g}$ could be a fraction in the interval (-1,1), which is not binary.

## 2 Problem 2.14

(a) A hypothesis set $H_i$ with finite VC dimension $d_{vc}$ means that $H_i$ can shatter at most $d_{vc}$ points and $H_i$ can have a most $2^{d_{vc}}$ dichotomies. So, $H = H_1 \cup ... \cup H_k$ can have at most $K \times 2^{d_{vc}}$ dichotomies. Now, $K \times 2^{d_{vc}} \leq 2^{K \times d_{vc}}$.
So, $d_{VC}(H) \leq 2^{K \times d_{vc}} < 2^{K \times (d_{vc}+1)}$

(b) Given, $2^l > 2Kl^{d_{vc}}$
For l datapoints, a hypothesis set $H_i$ with $d_{vc}$ can give at most $l^{d_{vc}} + 1$ dichotomies. so, $m_{Hi}(l) \leq l^{d_{vc}} + 1$. Therefore, for $H = H_1 \cup ... \cup H_K$ we can write $m_H(l) \leq Kl^{d_{vc}} + K$. Since, l, $d_{vc}$ and K are positive, so we may assume $Kl^{d_{vc}} \geq K$. So,

$$m_H(l) \leq 2Kl^{d_{vc}} \tag{1}$$

Given, $2^l > 2Kl^{d_{vc}}$, so replacing in the above inequality (1) we may write,

$$m_H(l) < 2^l$$

So, by definition, $d_{vc}(H) \leq l$

(c) From (a), $d_{vc} < K(dvc + 1)$ and from (b) $dvc \leq l$
Let use assume that $l = 7(dvc + K)log_2(Kd_{vc})$
From (b) it is given that

$$2^l > 2Kl^{dvc}$$
$$2^{7(dvc+K)log_2(Kd_{vc})} > 2K[7(dvc + K)log_2(Kd_{vc})]^{dvc}$$
$$(Kdvc)^{7(dvc+K)} > 2K[7(dvc + K)log_2(Kd_{vc})]^{dvc}$$

Taking $log_2$ on both sides,

$$7d_{vc}log_2(Kd_{vc}) + 7Klog_2(Kd_{vc}) > 1 + log_2K + d_{vc}log_27 + d_{vc}log_2(dvc + K) + d_{vc}log_2log_2(Kdvc) \tag{2}$$

If we assume that $K > 1$, then we may write,

$$d_{vc}log_2(Kd_{vc}) > d_{vc}log_2log_2(Kd_{vc}) \tag{3}$$
$$4d_{vc}log_2(Kd_{vc}) > 1 \tag{4}$$
$$d_{vc}log_2(Kd_{vc}) > log_2(K) \tag{5}$$
$$d_{vc}log_2(Kd_{vc}) \geq d_{vc}log_2(dvc + K) \tag{6}$$

Also, $d_{vc}log_27 < d_{vc}log_28 = 3dvc < 7Klog_2(Kdvc)$
So,

$$7Klog_2(Kd_{vc}) > d_{vc}log_27 \tag{7}$$

Adding 3,4,5,6,7, inequalities, we get

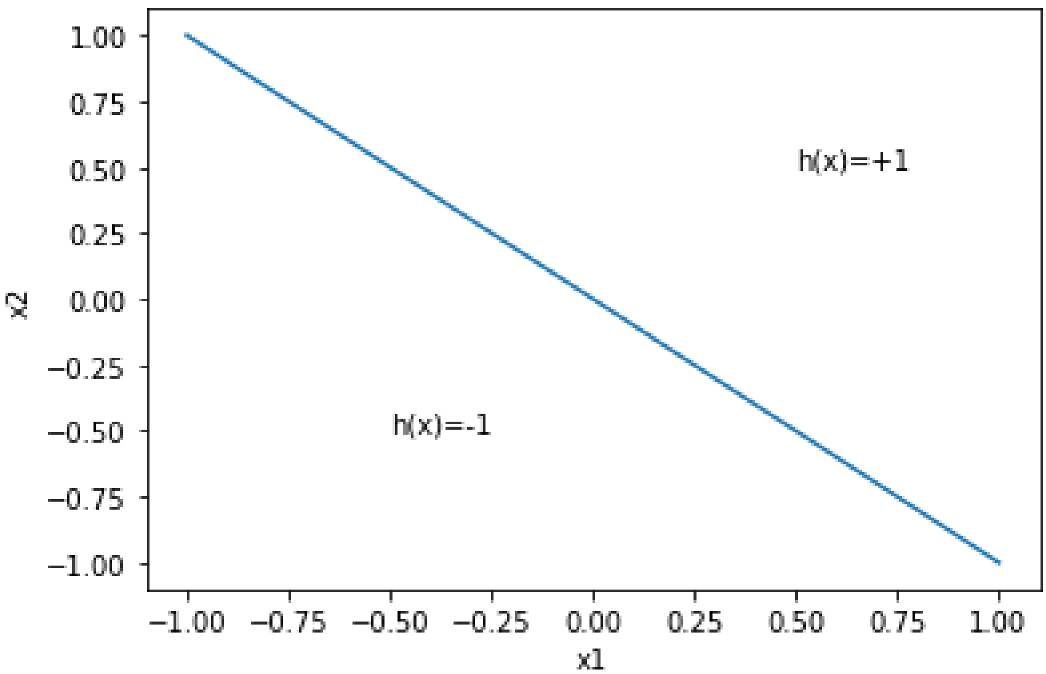
$$7d_{vc}log_2(Kd_{vc}) + 7Klog_2(Kd_{vc}) > 1 + log_2K + d_{vc}log_27 + d_{vc}log_2(dvc + K) + d_{vc}log_2log_2(Kdvc)$$

So, Inequality in (2) is true. So, $2^l > 2Kl^{dvc}$ holds with $l = 7(dvc + K)log_2(Kd_{vc})$

So, $d_{vc}(H) \leq min(K(dvc + 1), 7(dvc + K)log_2(Kd_{vc}))$

# 3 Problem 2.15

(a) An example of a monotonic classifier is shown below along with decision region where h(x)=sign(x1+x2):




(b) We can onsider a set of N points generated by first choosing one point and, then generating the next point by increasing the first component and decreasing the second component until N points are obtained. In this case for any datapoint, $\mathbf{x_i} > \mathbf{x_j}$ does not hold for any i,j. So, our hypothesis set H is able to generate any dichotomy on N points. So, $m_H(N) = 2^N$. So, $d_{vc} = \infty$

# 4 Problem 2.24

(a) From the data set D=$[(x_1, x_1^2), (x_2, x_2^2)]$, we can obtain the function:

$$\frac{g(x) - x1^2}{x1^2 - x2^2} = \frac{x - x1}{x1 - x2}$$
$$=> g(x) = \frac{(x - x1)(x1^2 - x2^2)}{x1 - x2} + x1^2$$
$$=> g(x) = (x1 + x2)x - x1x2$$

So, we can find average function,

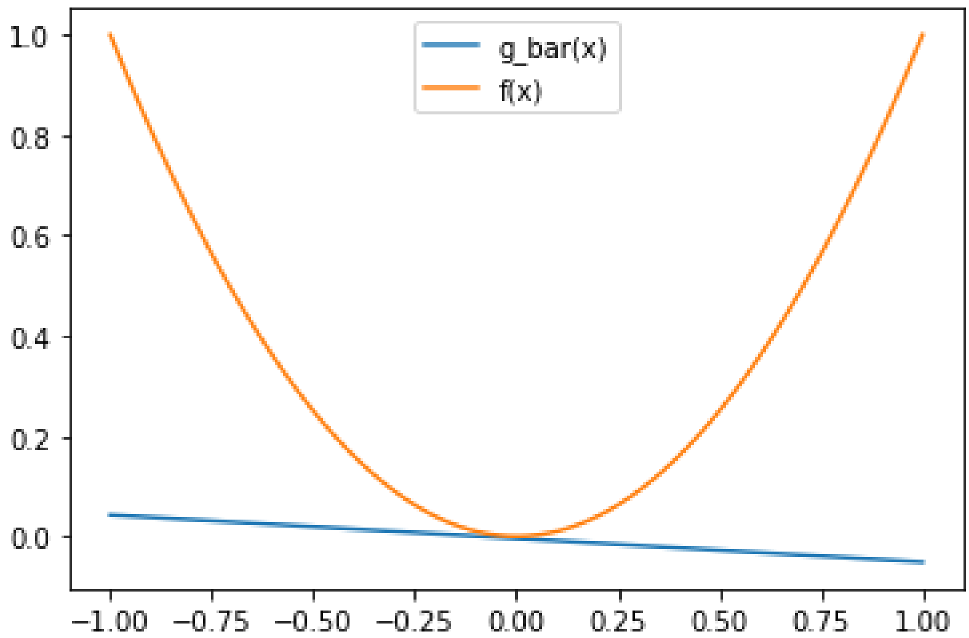
$$\bar{g}(x) = \frac{1}{2}\frac{1}{2}\int_{-1}^{1}\int_{-1}^{1}[(x_1 + x_2)x - x_1x_2]dx_1dx_2$$
$$= \frac{1}{2}\frac{1}{2}\int_{-1}^{1} x_2xdx_2$$
$$= 0$$

(b) We will generate 1000 numbers in range [-1,1] and fit f(x). Then for 1000 times, we choose two numbers x1 and x2 randomly from [-1,+1] again, and determine g(x) using the equation in (a) and $\{(x1, x1^2), (x2, x2^2)\}$
After getting 1000 g, we calculate $\bar{g}$ using previously generated 1000 g functions, and calculate variance using this $\bar{g}$ with respect to previously generated g functions.
We then calculate bias using calculated g and target function f(x). Finally, we compare $E[E_{out}]$ with bias + variance.

(c) After running the experiment, we get

$$\bar{g}(x) = -0.0469x - 0.0043$$

with bias= 0.2044, variance= 0.3324 and $E[E_{out}] = 0.5311$. Here $bias + variance = 0.5368 \approx E[E_{out}]$



(d) Here,

$$\begin{aligned} bias &= E[(\bar{g}(x) - f(x))^2] \\ &= \frac{1}{2}\int_{-1}^{1}(0 - x^2)^2 dx \\ &= \frac{1}{2}\int_{-1}^{1} x^4 dx \\ &= \frac{1}{5} \end{aligned}$$

$$\begin{aligned} variance &= E_x[E_D[(g^D(x) - \bar{g}(x))^2] \\ &= \frac{1}{2}\int_{-1}^{1}\frac{1}{2}\frac{1}{2}\int_{-1}^{1}\int_{-1}^{1}((x1 + x2)x - x1x2)^2 dx1dx2dx \\ &= \frac{1}{8}\int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1}[(x1 + x2)^2x^2 - 2(x1x2(x1 + x2))x + (x1x2)^2]dx1dx2dx \\ &= \frac{1}{8}\int_{-1}^{1}\int_{-1}^{1}\int_{-1}^{1}[(x1^2 + 2x1x2 + x2^2)x^2 - 2(x1^2x2 + x1x2^2)x + (x1x2)^2]dx1dx2dx \\ &= \frac{1}{8}\int_{-1}^{1}\int_{-1}^{1}[(\frac{2}{3} + x2^2)x^2 - \frac{4}{3}x2 + \frac{2}{3}x2^2]dx2dx \\ &= \frac{1}{8}\int_{-1}^{1}[(\frac{2}{3} + \frac{x2^3}{3})x^2 - \frac{4}{3}x2^2 + \frac{2}{9}x2^3|_{-1}^{1}]dx \\ &= \frac{1}{8}\int_{-1}^{1}[(\frac{2}{3} + \frac{2}{3})x^2 + \frac{4}{9}]dx \\ &= \frac{1}{8}(\frac{16}{9} + \frac{8}{9}) = \frac{1}{3} \end{aligned}$$

$E[E_{out} == bias + variance = \frac{1}{5} + \frac{1}{3} = \frac{8}{15}$