CS6350 - Homework/Assignment-4

Arnab Das(u1014840)

October 31, 2016

1: PAC-Learning

(1.a) Rule-1 states that you are free to combine any of the parts as they are. Given, there are N available parts, then for every part there are two options - either the part is chosen or it is discarded. It is analogous to a N-bit vector, where each part corresponds to a bit in the vector. The decision of choosing the part can be thought of as setting 1 to that bit position while the choice of discarding a part is equivalent to seeting 0 to that bit position. For such a set-up the size of the hypotheses space is equal to the possible numbers that the n-bit vector can accomodate. Hence,

Size of hypotheses space by Rule-1 = 2^N (Answer).

(1.b) Here both Rule-1 and Rule-2 are followed. Rule-2 suggests that the original parts are allowed to be broken into two distinct pieces before using. Thus, in combination of Rule-1 and Rule-2, there will be now four choices for every part, that is , the part is not chosen, or the part is chosen as original without cut, or the part is cut and the first piece is chosen, or the part is cut and the second piece is chosen. For the analogy of the n-bit vector, each bit can now take 4 values instead of just two. Hence,

Size of hypotheses space by Rule-1 and 2 = 4^N (Answer).

(1.c) Number of available parts = 6. Hence, N = 6.

Size of hypotheses space = $|H| = 4^6$

The allowed error, $\epsilon = 0.01$

Given, the probability with which the robot wants to predict correctly within ϵ is 99%. Then $(1 - \delta) = 0.99$.

Then, δ = 1 - 0.99 = 0.01

Note that, in this case the learning is not agnostic, since the true concept is contained in the hypotheses space. Hence, we need to use the non-agnostic learning formula for finding the lower bound on the number of examples:

$$m > \frac{1}{\epsilon}\left(ln(|H|) + ln(\frac{1}{\delta})\right) \tag{1}$$

Plugging in the values for $|H|, \epsilon, \delta$ we get the value of m as:

$$m > 1292.293$$

Rounding off to the ceil:

$$m > 1293$$

(Answer)

(2) Suppose H denotes the hypotheses space. Consider a hypothesis h $\in$ H. Let us define a random Variable 'X' , such that:

$$X = 0; h(x) = c(x)$$
$$X = 1; h(x) \neq c(x)$$

where x belongs to the distribution D over which the instance space is defined, and c the target concept.
If we select 'm' random independent examples from the distribution D and the mark the outcomes as $X_1, X_2, \ldots, X_m$, then $\frac{\sum X_i}{m}$ denotes the error fraction over the sample space. The **training error**, $error_s(h)$, is defines as the fraction of the training examples misclassified by the hypothesis h. Hence,

$$error_s(h) = \frac{\sum X_i}{m} \tag{2}$$

The true error, or generalization error, $error_D$ over the entire distribution D from which the examples are randomly drawn is defined as

$$error_D(h) \equiv \underset{x \in D}{P}[c(x) \neq h(x)]$$

The expected value of a discrete random variable, is the probability weighted average of all possible values, that is over the entire distribution. Hence, the from the above definition of $error_D$, we can say it is the expected value of $error_s$, the training error.

For PAC learnability, we would like that the true error is not worse than ϵ plus the training error. Then we can characterize the event $(error_D - error_s > \epsilon)$ as a bad event and would like the probability of such an event be upperbounded by a small probability δ. We can write the probability of the above event as following:

$$P[error_D > error_s + \epsilon]$$

In this problem, we are given that the error term is a multiplicative terms relative to the training error, that is true error is no worse than $(1+\epsilon)error_s$. Then we can write the above as:

$$P[error_D - (1+\epsilon)error_s > 0]$$

Rearranging the terms:

$$P[error_D > (1+\epsilon)error_s]$$

$$P[\frac{1}{1+\epsilon}error_D > error_s]$$

$$P\left[error_D\left(1 - \frac{\epsilon}{1+\epsilon}\right) > error_s\right] \tag{3}$$

Recalling the Chernoff bound, where for $X_1, X_2, \ldots, X_m$ being the outcomes of m independent trials, and the probability of a $P[X_i = 1] = p$ and $P[X_i = 0] = (1-p)$, then the expectation $E\left[\frac{\sum X_i}{m}\right]$=p, and the chernoff bound governs the probability that $\frac{\sum X_i}{m}$ will differ from p by some factor $0 < \gamma < 1$, as:

$$P\left[\frac{\sum X_i}{m} > (1+\gamma)p\right] \leq \exp(\frac{-mp\gamma^2}{3})$$

$$P\left[\frac{\sum X_i}{m} < (1-\gamma)p\right] \leq \exp(\frac{-mp\gamma^2}{2})$$

Recalling equation(2) and the definition of true error, we see that equation(3) is in exact formation of the second chernoff bound described above such that
$p = error_D$ and $\gamma = \frac{\epsilon}{1+\epsilon}$. Since $0 < \epsilon < 1$, then $0 < \gamma < 1$, which satisfies the condition on the form factor of the Chernoff bound. Thus, using the second chernoff bound in our problem, we get:

$$P\left[error_D\left(1 - \frac{\epsilon}{1+\epsilon}\right) > error_s\right] \leq \exp(\frac{-m.error_D.\left(\frac{\epsilon}{1+\epsilon}\right)^2}{2}) \tag{4}$$

The above was determined for a single hypethesis $h \in H$. For the entire hypotheses space the above equation has to be adjusted to :

$$P\left[error_D\left(1 - \frac{\epsilon}{1+\epsilon}\right) > error_s\right] \leq |H| \exp\left(\frac{-m.error_D.\left(\frac{\epsilon}{1+\epsilon}\right)^2}{2}\right) \tag{5}$$

We use the result of equation(5) to determine the number of training examples required to reduce this probability of failure below a desired level of δ. Hence:

$$|H| \exp\left(\frac{-m.error_D.\left(\frac{\epsilon}{1+\epsilon}\right)^2}{2}\right) \leq \delta$$

Taking natural logarithms on both sides and reaarnging the terms to single out m(number of examples for training), we get:

$$m \geq \frac{2(1+\epsilon)^2}{\epsilon^2.error_D}\left(ln|H| + ln|\frac{1}{\delta}|\right) \tag{6}$$

(Answer).

2: VC Dimensions

---

(1) Suppose we have a finite hypotheses space C. Let the instance space be 'X" and the VC(C) defined over the space 'X" is n. This implies that C is able to shatter atmost a subset of n instances out of 'X'. To shatter n instances, the number oof hypotheses required will be atleast 2^n. Hence,

$$|C| \geq 2^n$$
$$=> \log_2 |C| \geq n \times \log_2 2$$
$$=> n \leq \log_2 |C|$$

Hence, VC(C) $\leq \log_2 |C|$ (Proved).

(2.a) $H^X_{=k} = h \in 0, 1^X : |x : h(x) = 1| = k$ that is, the set of all functions that assign the value 1 to exactly k elements. Consider 1 sample point. The adversary can mark this point as a '1' label or a '0'. The class defintiion says that for a given fixed 'k', the hypotheses contained in this space are the ones that mark exactly k points as 1. Suppose k=1. The possible dichotomies of a single point is two, that is either the point is labeled as 1 or labeled as 0. Then, for k=1, it works for the dichotomy where the label is 1, but for label 0, it does not have a hypotheses in this class. The class for k=0, has a hypotheses for label=0, but it doesn't havs a hypotheses for the label being 1. Thus, for a specific fixed k, it cannot shatter a single point. Hence, **VC dimension of this class is 0**. (answer)

(2.b) $H^X_{\leq k} = h \in 0, 1^X : |x : h(x) = 1| \leq k or |x : h(x) = 0| \leq k$
That is a class of functions for a given fixed k, there exists hypothesis that can label atmost k elements as 1 or a 0. Let us consider a few rudimentary cases to identify the formal pattern:
case-1: single point - If this point is labeled 0, then there exists a $h \in H_1$ that can label a single point as 0, due to the fact $|x : h(x) = 0| \leq 1$, and holds true for all $k \geq 1$.
If this point is labeled 1, then also there exists a $h \in H_1$ that can label a single point as 1 due to the fact $|x : h(x) = 1| \leq 1$, and holds true for all $k \geq 1$.
case-2: 2 points - Consider 2 points. 2 points can have the following 4 labels: (0,0), (0,1), (1,0) and (1,1). In all 4 cases the $|x = 1|$ is atmost 2, and hence satisfied by $h \in H_2 : |x : h(x) = 1| \leq 2$ and also $|x = 0|$ is atmost 2 and hence satisfied by $h \in H_2 : |x : h(x) = 0| \leq 2$.
As the pattern is very evident that if the number of points, n, is less than or equal to k, then $H^X_{\leq k}$ can shatter the points.
Now let us consider, that the number of points n is greater than k, say k+1, for a given fixed k. There exists a labelling where all the points are labelled as 1, such that $|x = 1|$ = k+1. But the hypotheses class we have has hypothesis that can label **atmost** k points as 1. Hence, in this case we have a breakpoint and so the

(k+1) points cannot be shattered.
Hence, **VC dimension is k**. (answer)

(3)

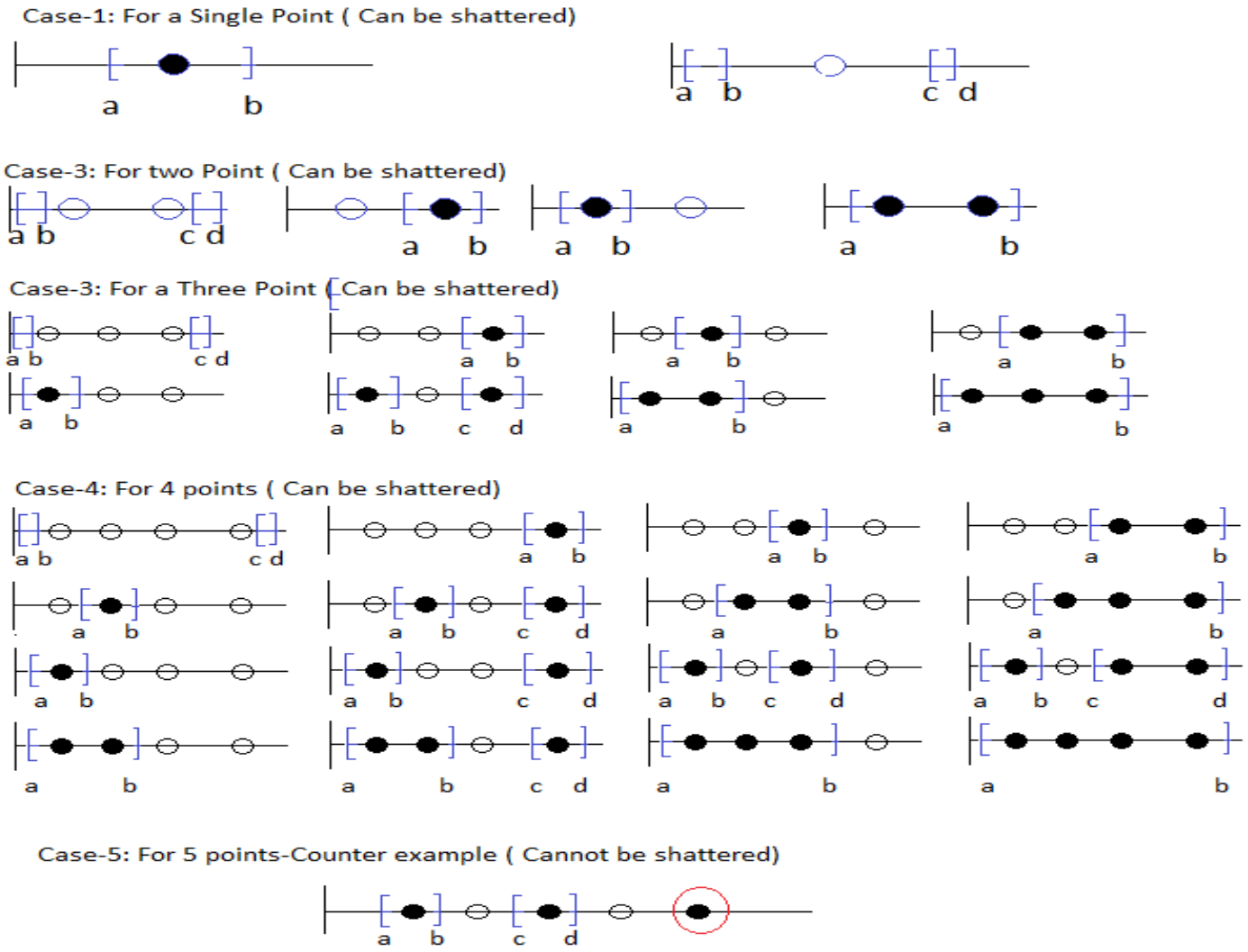
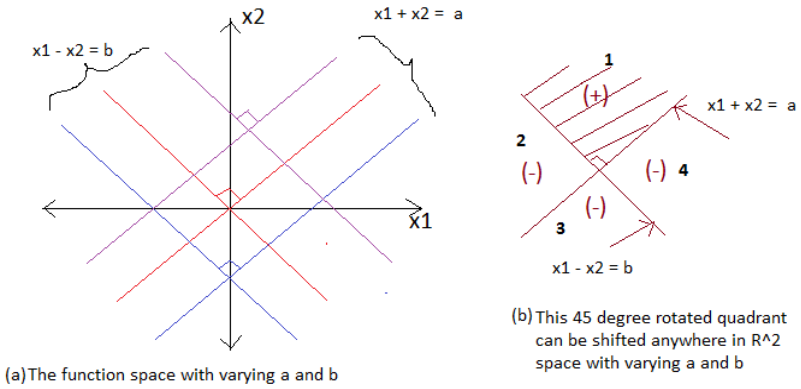


Figure 1: Shattering of real number instance space by two disjoint intervals

Instance space consisting of real numbers and a hypothesis space H consisting of two disjoint intervals, defined by $[a, b]$ and $[c, d]$.
In figure-1, we describe the possible shatterings of set of points on the real number line. For 5 points we can show a labelling, where two disjoint intervals cannot shatter the set of points.Hence, **VC dimension is 4** . (Answer).

(4) Each example point in R^2 . A function $h \in H$, where H is the concept class, is specified by 2 parameters a and b. An example x_1, x_2 is labeled '+' if and only if $x_1 + x_2 \geq a$ and $x_1 - x_2 \leq b$, else labeled as '-'. So, it looks like the figure-2,



(a)The function space with varying a and b

(b) This 45 degree rotated quadrant can be shifted anywhere in R^2 space with varying a and b

Figure 2: Description of the function space

Fig-2(a), shows how the h changes with varying a and b. Note that $x_1 + x_2 = a$ and $x_1 - x_2 = b$ are perpendicular to each other. Hence, as we vary a and b, the lines do not rotate, but only move in parallel. Thus, all the functions correspond to fig2(b), that is a 45 degree axes rotated 4 quadrant structure, and that can be formed in any part of R^2 , with varying a and b. We have the quadrants as 1,2,3,4 for ease of explanation later.
Now, let us consider case by case for varying set of points.
Case-1: Single Point - A single point is fairly obvious that it can be shattered. If it is labeled '-', then we can always find pair of (a,b) such that the point falls in one of 2,3,4 quadrants. If it is marked as '+', then as well we can find a pair (a,b) such that the point falls in the 1 quadrant.
Case-2: 2 points - Figure(3) shows the set of two points that can be shattered with this hypotheses space.
Fig(3)a shows for two points labeled (+,+) we can have a pair of (a,b) such that they lie in quadrant-1 of the 45 degree rotated 4 quadrant structure. Fig-3(b) analyses the case where labeling is (+,-). We show how the lines of 3(a) can be parallel shifted as the **blue** lines, to satisfy this labeling. Similarly, we show for the labelings of (-,+) and (-,-) in 3(c) and 3(d) respectively. Thus 2 points can be shattereds for this class.

Case-3: 3 points - Here, we strive to explain the different geomtries in which the points can be placed from the perspective of our hypothesis space. Figure(4) shows the different geomtries. The geometries that are symmetrical with respect to our hypetheses space are grouped together and only one of them will be considered.

In Figure(4) we show the counter-examples for all these geometries.
For **geometry-1**, since the '-' labeled point is above the two '+' labeled points and midway along x1 for the '+' labeled points, hence the given labeling for geometry-1 is not satisfiable by any $h \in H$.
For **geometry-2**, For this setting we show two different labelings that cannot be simultaneously true. Since, every hypothesos with (a,b) are correspondingly parallel, this given pair of labelling violates it.
For **geometry-3**, Similarly, here also we show two labelling , that will conflict.

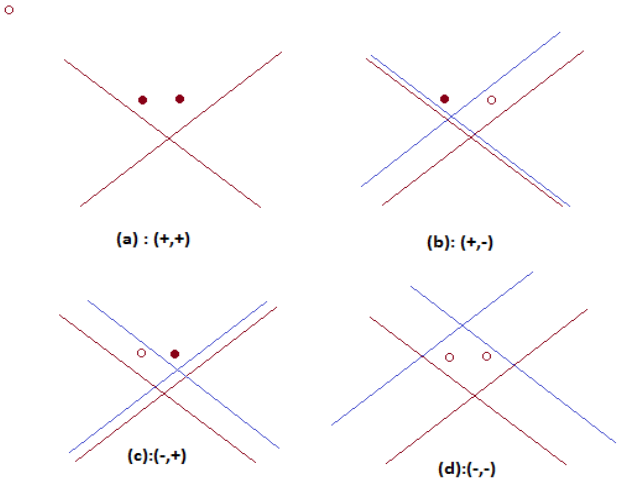


Figure 3: Shattering two points

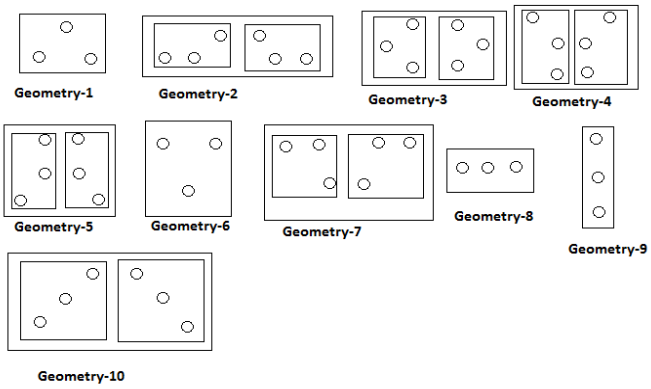


Figure 4: Possible geometries of 3 points

For **geometry-4,5,6,7**, For geometry 4,5,6,7, we show labellings, which does not have a feasible hypothesis, since the convex region of '+' will contain the '-' labeled point, contradicting the hypotheses space definition. For **geometry-8,9,10**, Geometry-8,9,10 covers the colinear points where again we again give counter examples labeling that is not satisfiable by any of the hypotheses. Basically, for colinear points, if the non-adjacent points are oppositely labeled, then it is not satisfiable.

Since, we get breakpoint for 3 points, hence **VC dimension is 2**.(Answer)

(5) Let two hypothesis classes H_1 and H_2 satisfy $H_1 \subseteq H_2$.
Suppose we find a set of instances of size d_1 that H_1 is able to shatter and a size $d_2 = d_1 + 1$, that H_1 cannot shatter. Then

$$d1 \leq VC(H1) < d2$$

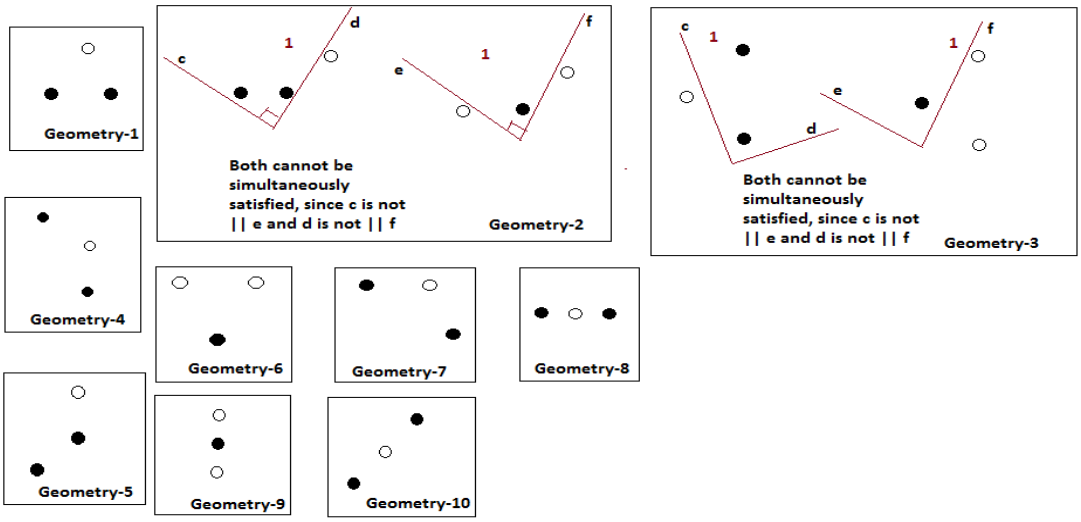


Figure 5: Counter Examples for each geometry of 3 points

Hence,

$$VC(H1) = d1$$

Since , $H_1 \subseteq H_2$, we can write: $H_2 = H_1 \cup \delta h$.
Since, H_2 contains all the hypotheses of H_1, so H_2 can **atleast** shatter the subset of instances that H_1 was able to shatter. Hence, for d1 set of instances, H_2 will be able to shatter using the same set of hypotheses H_1 had to shatter d_1. Hence:

$$VC(H_2) \geq d1 = VC(H_1)$$
$$VC(H_2) \geq VC(H_1) \tag{7}$$

(Proved).

3: AdaBoost

Using the formulae:
(1) $\epsilon_t = \frac{1}{2} - \frac{1}{2}(\sum_{i=1}^{m} D_t(i) y_i h(x_i))$
(2) $D_{t+1}(i) = \frac{D_t(i)}{z_t} \exp(-\alpha_t y_i h_t(x_i))$
(3) $\alpha_t = ln(\frac{1-\epsilon_t}{\epsilon_t})$
Given for first iteration:
h = $h_a(x) = \text{sgn}(x_1)$, $\epsilon_1 = 1/4$, $\alpha_1 = \frac{ln3}{2}$
Initialize $D_1 = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4},\right)$

Evaluating $D_2'(without - normalization) = \left(\frac{\sqrt[2]{3}}{4}, \frac{1}{4\sqrt[2]{3}}, \frac{1}{4\sqrt[2]{3}}, \frac{1}{4\sqrt[2]{3}}\right)$

$Z_1 = \sum_{i=1}^{4} D_2'(i) = \frac{\sqrt[2]{3}}{2}$

Hence, Normalized $D_2 = \left(\frac{1}{2}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right)$

Starting second iteration

Evaluating ϵ corresponding to the remaining hypotheses

$\epsilon_{2,h_b} = \frac{1}{6}$

$\epsilon_{2,h_c} = \frac{1}{2}$

$\epsilon_{2,h_d} = \frac{1}{6}$

Since both ϵ_{2,h_b} and ϵ_{2,h_d} are lowest and better than chance, we can pick any of h_b or h_d . So Choosing h_b for round 2, we evaluate $\alpha_2 = \frac{ln5}{2}$

Hence the data for Round-2 is:

h = h_b = sgn(x-2) ; $\epsilon_2 = \epsilon_{2,h_b} = \frac{1}{6}$, $D_2 = \left(\frac{1}{2}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right)$

Evaluating $D_3'(without - normalization) = \left(\frac{1}{2\sqrt[2]{5}}, \frac{\sqrt[2]{5}}{6}, \frac{1}{6\sqrt[2]{5}}, \frac{1}{6\sqrt[2]{5}}\right)$

$Z_2 = \frac{\sqrt[2]{5}}{3}$

Hence, Normalized $D_3 = \left(\frac{3}{10}, \frac{1}{2}, \frac{1}{10}, \frac{1}{10}\right)$

Starting third iteration

Evaluating ϵ corresponding to the remaining hypotheses

$\epsilon_{3,h_c} = \frac{7}{10}$

$\epsilon_{3,h_d} = \frac{1}{10}$

Since ϵ_{3,h_d} is lowest and better than chance, hence choosing h_d for round 3, and evaluating $\alpha_3 = ln(3)$

Hence, the data for Round-3 is:

h = $h_d = -sgn(x_2)$; $\epsilon_3 = \epsilon_{3,h_d} = \frac{1}{10}$, $D_3 = \left(\frac{3}{10}, \frac{1}{2}, \frac{1}{10}, \frac{1}{10}\right)$

Evaluating $D_4'(without - normalization) = \left(\frac{1}{10}, \frac{1}{6}, \frac{3}{10}, \frac{1}{30}\right)$

$Z_3 = \frac{3}{5}$

Hence, Normalized $D_4 = \left(\frac{1}{6}, \frac{5}{18}, \frac{1}{2}, \frac{1}{18}\right)$

Starting fourth iteration

Evaluating ϵ corresponding to the remaining hypotheses

$\epsilon_{4,h_c} = \frac{5}{6}$

Since, ϵ_{4,h_c} is worse than chance, we **discard this hypotheses** h_d.

Hence, our final combined function is :

$$H_{final} = sgn((ln(\sqrt[2]{3}))sgn(x_1) + (ln(\sqrt[2]{5}))sgn(x-2) - (ln3)sgn(x_2)) \tag{8}$$

The final table is given in Table-2:

Table 1: Table-Round1: $h_a(x), \epsilon_1 = \frac{1}{4}, \alpha_1 = \frac{ln3}{2}, Z_1 = \frac{\sqrt[2]{3}}{2}$

$X=[x_1,x_2]$	y_i	D_1	$D_1(i)y_ih_t(x_i)$	D_2
(1,-1)	-1	$\frac{1}{4}$	$-\frac{1}{4}$	$\frac{1}{2}$
(1,-1)	+1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{6}$
(-1,-1)	-1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{6}$
(-1,1)	-1	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{6}$

Table 2: Table-Round2: $h_b(x), \epsilon_2 = \frac{1}{6}, \alpha_2 = \frac{ln5}{2}, Z_2 = \frac{\sqrt[2]{5}}{3}$

$X=[x_1,x_2]$	y_i	D_2	$D_2(i)y_ih_t(x_i)$	D_3
(1,-1)	-1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{10}$
(1,-1)	+1	$\frac{1}{6}$	$\frac{-1}{6}$	$\frac{1}{2}$
(-1,-1)	-1	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{10}$
(-1,1)	-1	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{10}$

Table 3: Table-Round3: $h_d(x), \epsilon_3 = \frac{1}{10}, \alpha_3 = ln3, Z_3 = \frac{3}{5}$

$X=[x_1,x_2]$	y_i	D_3	$D_3(i)y_ih_t(x_i)$	D_4
(1,-1)	-1	$\frac{3}{10}$	$\frac{3}{10}$	$\frac{1}{6}$
(1,-1)	+1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{5}{18}$
(-1,-1)	-1	$\frac{1}{10}$	$\frac{-1}{10}$	$\frac{1}{2}$
(-1,1)	-1	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{18}$

$\epsilon_{4,h_c} = \frac{5}{6}$, worse than chance, hence discarded.

Final Hypothesis, $H_{final}(x) = sgn((ln(\sqrt[2]{3}))sgn(x_1) + (ln(\sqrt[2]{5}))sgn(x-2) - (ln3)sgn(x_2))$ (Answer).