Amir Kaplan – 208789172 , Daniel Alfasi - 318601622

Solution 1.a:

$$Denote\ K_1(x,y) = \varphi_1(\mathrm{x})^{\mathrm{T}} \cdot \varphi_1(y) = [\varphi_1(\mathrm{x})_1,\quad \varphi_1(\mathrm{x})_2, \dots,\quad \varphi_1(\mathrm{x})_N] \begin{bmatrix} \varphi_1(\mathrm{y})_1 \\ \vdots \\ \varphi_1(\mathrm{y})_N \end{bmatrix}$$

$$= \varphi_1(\mathrm{x})_1 \cdot \varphi_1(\mathrm{y})_1 + \varphi_1(\mathrm{x})_2 \cdot \varphi_1(\mathrm{y})_2 + \dots + \varphi_1(\mathrm{x})_N \cdot \varphi_1(\mathrm{y})_N = S_1$$

$$Denote\ K_2(x,y) = \varphi_2(\mathrm{x})^{\mathrm{T}} \cdot \varphi_2(y) = [\varphi_2(\mathrm{x})_1,\quad \varphi_2(\mathrm{x})_2, \dots,\quad \varphi_2(\mathrm{x})_K] \begin{bmatrix} \varphi_2(\mathrm{y})_1 \\ \vdots \\ \varphi_2(\mathrm{y})_K \end{bmatrix}$$

$$= \varphi_2(\mathrm{x})_1 \cdot \varphi_2(\mathrm{y})_1 + \varphi_2(\mathrm{x})_2 \cdot \varphi_2(\mathrm{y})_2 + \dots + \varphi_2(\mathrm{x})_K \cdot \varphi_2(\mathrm{y})_K = S_2$$

$Now\ we\ want\ to\ show\ that:$
$K(x,y) = \varphi(x)^T \cdot \varphi(y) = 5K_1(x,y) + 4K_2(x,y) = 5S_1 + 4S_2$

$We\ define\ the\ following\ mapping - \varphi: \mathbb{R}^n \to \mathbb{R}^{N+K}\ s.t$
$\varphi(x) = [\sqrt{5}\varphi_1(x)_1,\quad \sqrt{5}\varphi_1(x)_2, \dots,\quad \sqrt{5}\varphi_1(x)_N,\ 2\varphi_2(x)_1,\quad 2\varphi_2(x)_2, \dots,\quad 2\varphi_2(x)_K,]$
$when\ \varphi_i(x) = [\varphi_\mathrm{i}(\mathrm{x})_1,\quad \varphi_\mathrm{i}(\mathrm{x})_2, \dots,\quad \varphi_\mathrm{i}(\mathrm{x})_{w_i}]\ and\ i \in \{1,2\}, w_1 = N, w_2 = K$
$Notice\ that\ it\ holds\ that:$

$K(x,y) = \varphi(x)^T \cdot \varphi(y)$
$= [\sqrt{5}\varphi_1(x)_1,\quad \sqrt{5}\varphi_1(x)_2, \dots,\quad \sqrt{5}\varphi_1(x)_N,\ 2\varphi_2(x)_1,\quad 2\varphi_2(x)_2, \dots,\quad 2\varphi_2(x)_K,] \cdot$
$[\sqrt{5}\varphi_1(y)_1,\quad \sqrt{5}\varphi_1(y)_2, \dots,\quad \sqrt{5}\varphi_1(y)_N,\ 2\varphi_2(y)_1,\quad 2\varphi_2(y)_2, \dots,\quad 2\varphi_2(y)_K,]^T$
$= 5S_1 + 4S_2 = 5K_1(x,y) + 4K_2(x,y)$ ■

Solution 1.b:

$Assume\ that\ \varphi_1: \mathbb{R}^n \to \mathbb{R}^m, \varphi_2: \mathbb{R}^n \to \mathbb{R}^K$

$Yes.first, since\ there\ exist\ a\ linear\ seperator\ fo\ the\ data\ under\ the\ mapping$
$\varphi_1: \mathbb{R}^n \to \mathbb{R}^m\ with\ the\ weight's\ vector\ w \in \mathbb{R}^{m+1}, it\ holds\ that:$

$$Sgn\left(w_0 + \varphi_1(x) \cdot \begin{bmatrix} \mathrm{w}_1 \\ \vdots \\ \mathrm{w}_\mathrm{m} \end{bmatrix}\right) is\ a\ linear\ classifier.$$

$We\ want\ to\ show\ that\ there\ exist\ w' \in \mathbb{R}^{m+K+1}\ s.t:$

$$Sgn\left(w_0{}' + \varphi(x) \cdot \begin{bmatrix} \mathrm{w}_1{}' \\ \vdots \\ \mathrm{w}_{\mathrm{m+K}}{}' \end{bmatrix}\right) = Sgn\left(w_0 + \varphi_1(x) \cdot \begin{bmatrix} \mathrm{w}_1 \\ \vdots \\ \mathrm{w}_\mathrm{m} \end{bmatrix}\right) \forall x \in D\ and\ then\ we\ could$$

$$conclude\ that\ w_0{}' + \varphi(x) \cdot \begin{bmatrix} \mathrm{w}_1{}' \\ \vdots \\ \mathrm{w}_{\mathrm{m+K}}{}' \end{bmatrix} is\ a\ linear\ seperator\ for\ the\ data\ under\ the$$

$mapping\ \varphi: \mathbb{R}^n \to \mathbb{R}^{m+K}\ with\ the\ weight's\ vector\ w'.$

$Define\ w' \in \mathbb{R}^{m+K+1}\ in\ the\ following\ order:$

$$w' = [\mathrm{w'}_0,\quad w'_1, \dots,\quad \mathrm{w'}_{\mathrm{m+K+1}}] = \left[w_0,\quad \frac{w_1}{\sqrt{5}}, \dots,\quad \frac{w_m}{\sqrt{5}},\ 0_{m+1},\quad 0_{m+2}, \dots,\quad 0_{m+K},\right]$$

$Hence:$

$$Sgn\left(w_0{}' + \varphi(x)\cdot\begin{bmatrix}\mathrm{w_1}'\\ \vdots \\ \mathrm{w_{m+K}}'\end{bmatrix}\right)$$

$$= Sgn(w_0 + [\sqrt{5}\varphi_1(x)_1,\quad \sqrt{5}\varphi_1(x)_2, \dots,\quad \sqrt{5}\varphi_1(x)_N,\ 2\varphi_2(x)_1,\quad 2\varphi_2(x)_2, \dots,\quad 2\varphi_2(x)_K,]\cdot \left[w_0,\quad \frac{w_1}{\sqrt{5}}, \dots,\quad \frac{w_m}{\sqrt{5}},\ 0_{m+1},\quad 0_{m+2}, \dots,\quad 0_{m+K},\right]^T)$$

$$= Sgn(w_0 + w_1\varphi_1(x)_1 + w_2\varphi_1(x)_2 + \cdots + w_m\varphi_1(x)_m)$$

$$= Sgn\left(w_0 + \varphi_1(x)\cdot\begin{bmatrix}\mathrm{w_1}\\ \vdots \\ \mathrm{w_m}\end{bmatrix}\right) \qquad \blacksquare$$

Solution 1.c:

$We\ define\ \varphi: S \to \mathbb{R}^N\ as\ follows:$

$$\varphi(x) = \left[\underbrace{3{,}3,\dots,}_{|x|}\underbrace{0{,}0,\dots,0}_{N-|x|}\right], \forall x \in S.$$

$i.e, \varphi\ is\ a\ vector\ in\ \mathbb{R}^N\ where\ the\ first\ |x|\ entries\ of\ it\ are\ 3\ and\ the\ last$
$N - |x|\ entries\ are\ 0.$

$By\ defining\ \varphi\ this\ way\ we\ can\ show\ that\ it\ satisfies\ the\ requrements:$

$Let\ x, y \in S\ for\ some\ S = \{1,2,\dots,N\}\ and\ w.l.o.g.\ assume\ |x| < |y|.\ it\ holds\ that:$

$$\varphi(x)\cdot\varphi(y) = \left[\underbrace{3{,}3,\dots,}_{|x|}\underbrace{0{,}0,\dots,0}_{N-|x|}\right]\cdot\left[\underbrace{3{,}3,\dots,}_{|y|}\underbrace{0{,}0,\dots,0}_{N-|y|}\right]^T$$

$$= \sum_i^{|x|} 3\cdot 3 + \sum_i^{|y|-|x|} 0\cdot 3 + \sum_i^{N-|y|} 0\cdot 0 = 9\min(x,y)$$

$Thus, by\ defining\ f(x,y) = \min(x,y)\ we\ get:$
$K(x,y) = 9f(x,y) = 9\min(x,y) = \varphi(x)\cdot\varphi(y)\ and\ so$
$K\ is\ a\ valid\ kernel\ with\ respect\ to\ \varphi \qquad \blacksquare$

Solution 2:

We want to find the maximum possible revenute under the constraint that out budget is $20{,}000\$$. *We therefore define*:

$B(h,s) = 20h + 170s - 20000$ *to be our constraint function, now, we can define the lagrangian* : $L(h,s,\lambda) = R(h,s) + \lambda \cdot B(h,s) = 200h^{\frac{2}{3}}s^{\frac{1}{3}} + \lambda(20h + 170s - 20000)$.

We saw that by maximizing such lagrangian we maximize $R(h,s)$ *and maintain our constraint, as the gradient of* $L(h,s,\lambda)$ *encapsulates both of our needs when it is equal to* $\vec{0}$.

This, we now turn to solve $\nabla L(h,s,\lambda) = \vec{0}$.

$$\frac{\partial L(h,s,\lambda)}{\partial h} = \frac{2}{3}\cdot 200s^{\frac{1}{3}}h^{-\frac{1}{3}} + 20\lambda = 0 \rightarrow \frac{400}{3}\cdot\left(\frac{s}{h}\right)^{\frac{1}{3}} = -20\lambda \rightarrow \left(\frac{s}{h}\right) = -\frac{3\lambda}{20}$$

$$\xrightarrow{t=\frac{s}{h}} t^{\frac{1}{3}} = -\frac{3\lambda}{20} \rightarrow t = -\frac{3\lambda}{20}\cdot t^{\frac{2}{3}} \qquad \boxed{1}$$

$$\frac{\partial L(h,s,\lambda)}{\partial s} = \frac{1}{3}\cdot 200\cdot s^{-\frac{2}{3}}\cdot h^{\frac{2}{3}} + 170\lambda = 0 \rightarrow \frac{200}{3}\cdot\left(\frac{h}{s}\right)^{\frac{2}{3}} = -170\lambda \rightarrow$$

$$\left(\frac{s}{h}\right)^{-\frac{2}{3}} = -\frac{51\lambda}{20} \xrightarrow{t=\frac{s}{h}} t^{\frac{-2}{3}} = -\frac{51\lambda}{20} \rightarrow 1 = -\frac{51\lambda}{20}\cdot t^{\frac{2}{3}} \rightarrow \frac{1}{17} = -\frac{3\lambda}{20}\cdot t^{\frac{2}{3}} \qquad \boxed{2}$$

From $\boxed{1}, \boxed{2}$ *we get* : $t = \frac{s}{h} = \frac{1}{17} \rightarrow h = 17s$, *we also have* :

$$\frac{\partial L(h,s,\lambda)}{\partial \lambda} = 20h + 170s - 20000 = 0 \text{ so for } h = 17s \text{ we get}$$

$$340s + 170s = 20000 \rightarrow 510s = 20000 \rightarrow s = \frac{2000}{51} \text{ amd so : } h = \frac{2000}{3}$$

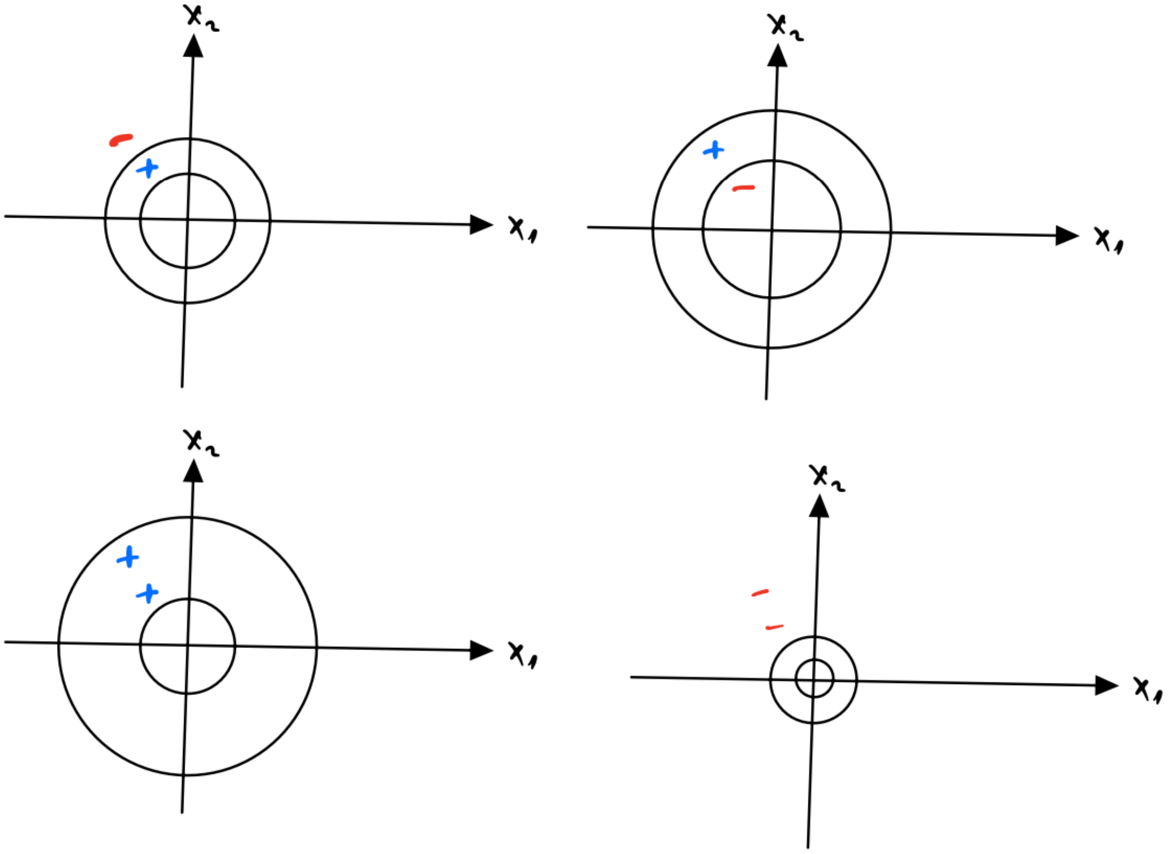
Therefore, for $(h,s) = \left(\frac{2000}{3}, \frac{2000}{51}\right)$

we get that our revenue is maximal. so, our maximal revenue given that our budget is $20000\$$ *is* :

$$R\left(\frac{2000}{3}, \frac{2000}{51}\right) = 200\cdot\left(\frac{2000}{3}\right)^{\frac{2}{3}}\cdot\left(\frac{2000}{51}\right)^{\frac{1}{3}} = 51854.81583 \qquad \blacksquare$$

Solution 3.a:

First, we show that $VC(H) \geq 2$:



So indeed, VC(H) ≥ 2.

Now, we proceed to show that $VC(H) < 3$.
we note that there are 2 *cases of convex hull*
for 3 *points in* $\mathbb{R}^2$ *which are : the* 3 *points are colinear, or the* 3 *points form a triangle*
for both cases, denote $r_{x_1}, r_{x_2}, r_{x_3}$ *to be the corresponding distances of points* x_1, x_2, x_3
from the origin and w.l.o.g. assume that $r_{x_1} \leq r_{x_2} \leq r_{x_3}$.
So, for the labeling : $x_1 = +, x_2 = -, x_3 = +$ *. it holds that*

$\forall h \in H \ \exists i \in \{1,2,3\} (h(x_i) \neq y_i)$, *Since out hypothesis space is all origin − centered rings it's impossible to exclude* x_2 *from within the ring without misclassifying one of the other instances, So ,* $VC(H) < 3$.
Since $VC(H) \geq 2$ *and* $VC(H) < 3$ *we dudce that* $VC(H) = 2$. ■

Solution 3.b:

We show that C is $PAC -$ learnable by L using H. in order to show that we will prove

that $\forall 0 < \epsilon < \frac{1}{2}, \forall 0 < \delta < \frac{1}{2}$, and for all $c \in C$ and distributions π over X, the following holds: with data drawn independently according to π, L will output, with probability at least $(1 - \delta)$, a hypothesis $h \in H$ such that $error_{\pi}(h) \le \epsilon$ as well as prove that L operates in time (and sample) complexity that is polynomial in $\frac{1}{\epsilon}$ and ,$\frac{1}{\delta}$.

Let $c \in C$, $\epsilon, \delta \in \left(0, \frac{1}{2}\right)$.

Denote r_1^*, r_2^* s.t $r_1^* \le r_2^*$ to be the radiuses of the target origin center annulus.

Assume that are our learner is consistent. which means that:

Given the training examples

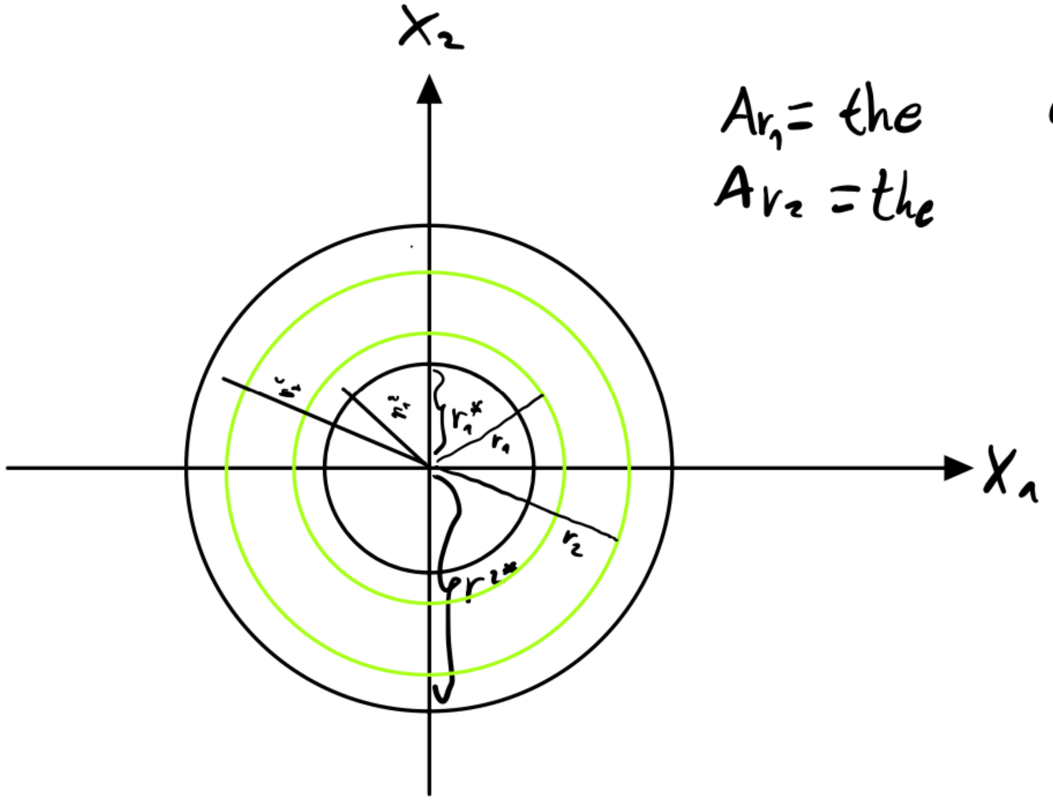
Find positive points with the maximum distance from the origin and minimum distance from the origin,

Define:

$r_1 =$ the distance of the positive point with the minimum distance from the origin,

$r_2 =$ the distance of the positive point with the maximum distance from the origin.

Output $h(r_1, r_2)$.



A_{r_1}= the area

A_{r_2} = the area

Consider training data, $D \in R^2$.
Case 1 :
Assume that D visits each one of the 2 *sets* A_i *defined above.* (*recall that* $\pi(A_i) = \frac{\epsilon}{2}$)
therefore our $error_\pi(x) \le \epsilon$.
Otherwise :
Every point in D never visited at least one of A_1, A_2.
therefore $P(\{D \in X^m: L(D), c) > \epsilon\}) \le \sum_{i=1}^{2}\left(P(X - A_i)\right)^m \le 2\left(1-\frac{\epsilon}{2}\right)^m \le \ 2e^{\left(-\frac{m\epsilon}{2}\right)}$
Hence :

$$\left(1-\frac{\epsilon}{2}\right)^m \le 2e^{\left(-\frac{m\epsilon}{2}\right)} \le \delta \to \ e^{-\frac{m\epsilon}{2}*\ln(e)} \le e^{\ln\left(\frac{\delta}{2}\right)} \to -\ln\left(\frac{\delta}{2}\right) \le \frac{m\epsilon}{2} * \ln(e) \to \ln\left(\frac{2}{\delta}\right) \le \frac{m\epsilon}{2}$$

$$\frac{2\ln\left(\frac{2}{\delta}\right)}{\epsilon} \le m$$

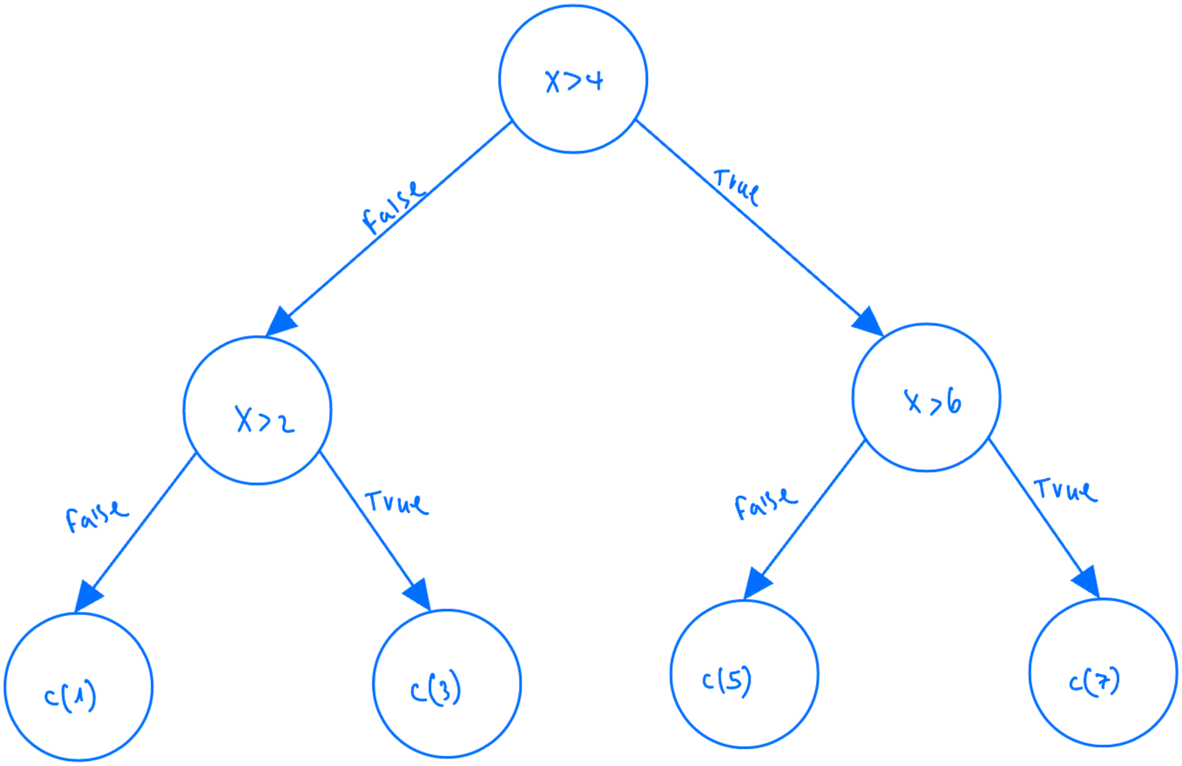
Solution 3c:

In order to get with 95% *confidence a hypothesis with at most* 5% *error we set* $\epsilon = 0.05$, $\delta = 0.05$.
Calculating the complexity with the bounds we found in the previous questions:
Sample complexity with the bound we found in 3.*b* : $m \ge 147.5$
Sample complexity with the bound for infinite H : $m \ge 2992.9$

We found that with the bound of 3.*b we have tighter bound as we learned in the lecture that the bound from a direct calculation encapsulates more information about the data since it assumes it comes from a certain distriubtion whereas the bound for infinite H does not assume anything about the data we are working on, and it only takes into account the VC dimensions of our H.*

Solution 4.a:

$\rightarrow VC(H_3) \geq 4$:
$We\ choose\ the\ following\ set\ of\ points\ \{1,3,5,7\}.$
$There\ will\ be\ a\ dichotomy\ for\ the\ set\ of\ points.$
$Denote\ the\ partition\ for\ the\ dichotomy\ as\ c(i), \forall i \in \{1,3,5,7\}$
$We\ will\ show\ that\ a\ binary\ decision\ tree\ with\ 7\ nodes\ (3\ internal\ nodes\ and\ 4\ leaf\ nodes)$
$can\ indeed\ shatter\ this\ set.$



$\forall c \in C\ (concept)\ we\ can\ always\ choose\ h\ s.t\ h(x_i) = c(x_i)\ , \forall x_i \in \{1,3,5,7\}.$

$\leftarrow VC(H_3) < 5$:
$Let's\ denote\ the\ set\ of\ 5\ points\ as\ X = \{x_1, x_2, x_3, x_4, x_5\}\ where\ x_1 < x_2 < x_3 < x_4 < x_5.$

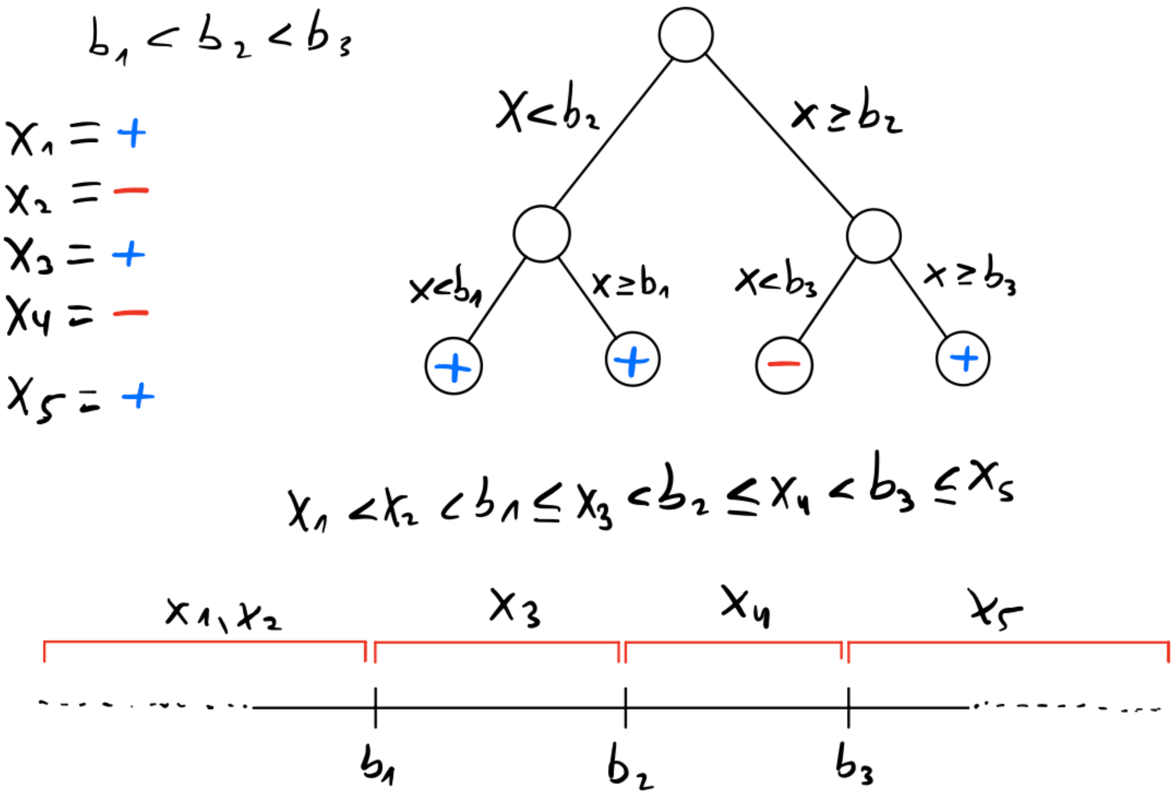
$We\ want\ to\ show\ that\ there\ exists\ a\ labeling\ of\ X\ that\ cannot\ be\ realized\ by\ a\ binary$
$decision\ tree\ with\ 7\ nodes\ (3\ internal\ nodes\ and\ 4\ leaf\ nodes).$

$Consider\ the\ labeling\ L = \{+, -, +, -, +\}.$

$A\ binary\ decision\ tree\ with\ 7\ nodes\ partitions\ the\ input\ space\ into\ 4\ regions, each$

corresponding to a leaf node. Each leaf node can output either $+$ or $-$
Let's denote the decision boundaries of the tree as b_1, b_2 and b_3, where $b_1 < b_2 < b_3$.
There boundaries partition the input space into the following 4 regions:

1. $R_1 = (-\infty, b_1)$
2. $R_2 = [b_1, b_2)$
3. $R_3 = [b_2, b_3)$
4. $R_4 = [b_3, \infty)$



Since we have 5 points and 4 regions, by the Pigeonhole Principle, there must be at least one region that contains at least two points. w.l.o.g., let's say x_1, x_2 fall into the same region, say R_1.

In the labeling L, x_1 and x_2 have different labels. However, all points in the same region get assigned the same label by the decision tree. Therefore, there is no way for the decision tree to realize the labeling L.

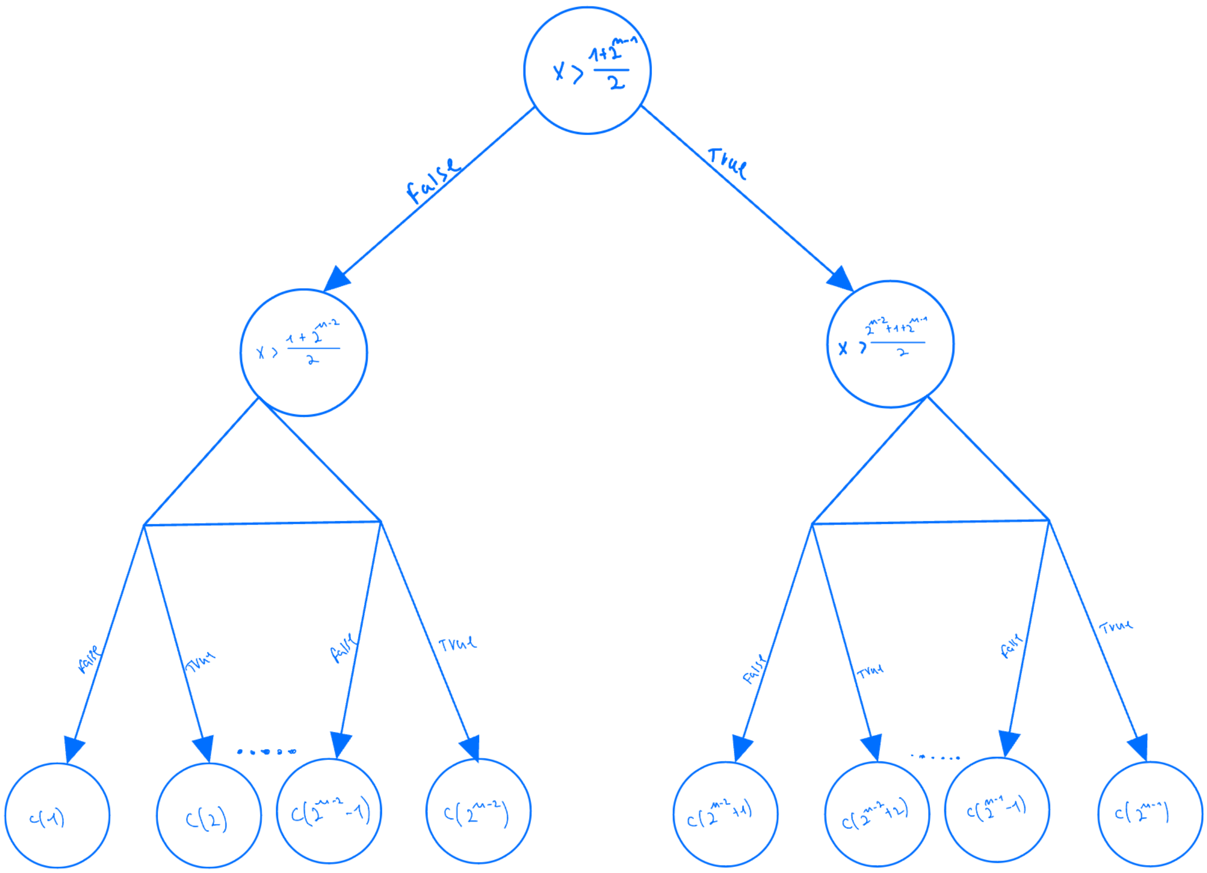
This shows that a binary decision tree with 7 nodes cannot shatter a set of 5 points, and hence the VC dimension of the hypothesis space of all x-node decision treewith $m \leq 3$ is less than 5.

Since we previously showed that this hypothesis space can shatter a set of 4 points, we conclude that the VC dimension of the hypothesis space of all "$x -$ node decision tree" with $m \leq 3$ is exactly 4

Solution 4.b:

$\rightarrow VC(H_m) \geq 2^{m-1}$:
We choose the following set of points $\{1, 2, \ldots, 2^{m-1}\}$.
Let be S be a dichotomy for the set of points.
Denote the partition for S as $c(i), \forall i \in \{1,2, \ldots, 2^{m-1}\}$
We will show that a binary decision tree with $2^n - 1$ nodes can indeed shatter this set.

We define the decision boundary for each $v_i \in V$(vertices of the decision tree) as follows : Denote $S_i = \{n_1, n_2, \ldots n_k\}$ and B_i to be the decision boundary question for each v_i.
$\forall v_i \in V$, B_i defined as $x > \frac{\min(S_i) \; + \max(S_i)}{2}$, if we went to the left child set S_i corresponding to v_i for the left portion of the set S_{i-1} (old S_i)
and if we went to the right child set S_{i+1} corresponding to v_{i+1} for the right portion of the set $S_{i-1.}$ Therefore we get the following diagram:



$\forall c \in C$ (*concept*) *we can always choose* h *s.t* $h(x_i) = c(x_i)$, $\forall x_i \in \{1,2 \ldots, 2^{m-1}\}$.

$\leftarrow VC(H_m) < 2^{m-1} + 1$:
Let's denote the set of $2^{m-1} + 1$ *points as* $X = \{x_1, x_2, \ldots, x_{2^{m-1}+1}\}$ *where*
$x_1 < x_2 < \cdots < x_{2^{m-1}+1}$.

We want to show that there exists a labeling of X *that cannot be realized by a binary decision tree with* $2^m - 1$ *nodes.*

Consider the labeling $L = \{+, -, +, -, \ldots, (-1)^{2^{m-1}+1}\}$.

A binary decision tree with $2^m - 1$ *nodes partitions the input space into* 2^{m-1} *regions,*
each corresponding to a leaf node. Each leaf node can output either $+$ *or* $-$
Let's denote the decision boundaries of the tree as $b_1, b_2, \ldots, b_{2^{m-1}-1}$,
where $b_1 < b_2 < \cdots < b_{2^{m-1}-1}$.
There boundaries partition the input space into the following 2^{m-1} *regions*:

$R_1 = (-\infty, b_1)$
$R_2 = [b_1, b_2)$
$\vdots$
$R_{2^{m-1}} = [b_{2^{m-1}-1}, \infty)$

Since we have $2^{m-1} + 1$ points and 2^{m-1} regions, by the Pigeonhole Principle, there must be at least one region that contains at least two points. w.l.o.g. let's say x_1, x_2 fall into the same region, say R_1.

In the labeling L, x_1 and x_2 have different labels. However, all points in the same region get assigned the same label by the decision tree. Therefore, there is no way for the decision tree to realize the labeling L.

This shows that a binary decision tree with $2^m - 1$ nodes cannot shatter a set of $2^{m-1} + 1$ points, and hence the VC dimension of the hypothesis space of all x-node decision tree with m is less than $2^{m-1} + 1$.

Since we previously showed that this hypothesis space can shatter a set of 2^{m-1} points, we conclude that the VC dimension of the hypothesis space of all "$x -$ node decision tree" with m is exactly 2^{m-1}