# Advanced Machine Learning

Sebastian Cojocariu - Group 407 IA

April 2021

## 1 Exercise 1

Give an example of a finite hypothesis class H with VCdim(H) = 2021. Justify your choice.

Solution:

We will prove a more general fact, namely that the infinite class of all homogeneous linear halfspaces classifiers in $\mathbb{R}^d$ has a VC dimension of $d$. Denote this class by $H_d,\ d \in \mathbb{N}_{>0}$.

$H_d = \{h_w : \mathbb{R}^d \to \{-1, 1\} | w \in \mathbb{R}^d\}$, where $h_w(x) = \begin{cases} 1, \text{ if } w^T * x \geq 0 \\ -1, \text{ if } w^T * x < 0 \end{cases}$

We will then choose only a finite subset from $H_{2021}$ that contains exactly $2^{2021}$ functions from $H$ to solve the exercise.

There are 2 steps to complete the proof:

i) Find a set of $d$ points in $\mathbb{R}^d$ that can be shattered by $H_d$ . We can choose the set $S$ of standard unit vectors $S = \{e_0, ..., e_{d-1}\}$. Arranging these points in a matrix form (each row represents the coordinates of a point), call it $A$, we have that $A = I_{2021}$, so we need to find $2^d$ column vectors $Y = \{y_0, ..., y_{2^d-1}\}$ such that each entry is either positive or negative (it does not matter which numbers are as long as these vectors are not equals when applying the sign function: $\text{sign}(x) = \begin{cases} 1, \text{ if } x \geq 0 \\ -1, \text{ if } x < 0 \end{cases}$), as well as $2^d$ column vectors $w_0, ..., w_{2^d-1} \in \mathbb{R}^d$ such that $Aw_i = y_i, i \in \{0, \ldots, 2^d - 1\}$. But since $A = I_{2021}$ and $h_w(A) = \text{sign}(Aw) \to h_w(A) = \text{sign}(w)$, for all w $\in \mathbb{R}^d$, we can just choose $w_i = y_i \in \{-1, 1\}^d$. This shows that these $d$ points are shattered by $H_d$.

ii) Prove that no set of $d + 1$ points in $\mathbb{R}^d$ can be shattered by $H_d$. Again, order each point as a row in a matrix called A. Since there are $d + 1$, it means that at least one row is a linear combination of the others. WLOG, assume $r_d = a_0 r_0 + ... + a_{d-1} r_{d-1}, a_i \in \mathbb{R}$, with $|a_0| + ... + |a_{d-1}| \neq 0$. Then, $Aw = \begin{bmatrix} r_0 w \\ r_1 w \\ \vdots \\ r_{d-1} w \\ a_0 r_0 w + ... + a_{d-1} r_{d-1} w \end{bmatrix}$. In this case, the labeling $Y = \begin{bmatrix} \text{sign}(a_0) \\ \text{sign}(a_1) \\ \vdots \\ \text{sign}(a_{d-1}) \\ -1 \end{bmatrix}$ would not be possible, since the first $d$ conditions from $Y$ imply that $a_0 r_0 w + ... + a_{d-1} r_{d-1} w \geq 0$, or equivalently $\text{sign}(a_0 r_0 w + ... + a_{d-1} r_{d-1} w) = 1$ (each term of the summation is positive). This means that the set cannot be shattered by $H$.

For our case: Since we are extracting a subset $H_{small}$ from the $H_{2021}$, we know that VCdim($H_{small}$) $\leq$ VCdim($H_{2021}$) $=$ 2021 (1). It remains to select 2021 points $S = \{e_0, ..., e_{2020}\}$ consisting of the standard unit vectors and order them as before in a matrix $A = I_{2021}$ (each row represent a point from the set we are trying to shatter) and $H_{small} = \{h_w \in H_{2021} | w \in \{\begin{bmatrix} \pm 1 \\ \pm 1 \\ \vdots \\ \pm 1 \end{bmatrix}\} = \{-1, 1\}^{2021}\}$ (finite, since it has exactly $2^{2021}$ classifiers). Since the set of possible labels is $L = \{\begin{bmatrix} \pm 1 \\ \pm 1 \\ \vdots \\ \pm 1 \end{bmatrix}\}$, $A = I_{2021}$ and $h_w(A) = \text{sign}(Aw) \rightarrow h_w(A) = \text{sign}(w)$, for all w $\in \mathbb{R}^d$, and w ranges from $\{-1, 1\}^{2021}$ we obtain $H_{small}$ shatters the points that A represents, resulting that VCdim($H_{small}$) $\geq$ 2021 (2). Combining (1) and (2) gives us VCdim($H_{small}$) $=$ 2021, as desired.

Another example: One could choose any finite set $|X| = 2021$ and define $H = \{f | f \in \{0, 1\}^X\}$ the set of all possible functions $f : X \rightarrow \{0, 1\}$ (obviously, $|H| = 2^X$ which is finite). Now, since $X$ is shattered by $H$ by definition (containing all the possible mappings), and since $X$ is the domain of these functions (no larger set can be chosen), it follows that VCdim($H$) $=$ 2021, as required. If we want the domain to be infinite as well, just add an infinite number of points and extend each $f \in H$ to label these new points as 0. Of course, $|X| = 2021$ is still shattered by H, but no set that contains any new points can be shattered (as there is no $f \in H$ that will label them as 1). Again, it follows that VCdim($H$) $=$ 2021.

# 2 Exercise 2

Consider $H_{balls}$ to be the set of all balls in $\mathbb{R}^2$: $H_{balls} = \{B(x, r), x \in \mathbb{R}^2, r \geq 0\}$, where $B(x, r) = \{y \in \mathbb{R}^2 | ||y - x||_2 \leq r\}$. As mentioned in the lecture, we can also view $H_{balls}$ as the set of indicator functions of the balls $B(x, r)$ in the plane: $H_{balls} = \{h_{x,r} : \mathbb{R}^2 \rightarrow \{0, 1\}, h_{x,r} = 1_{B(x,r)}, x \in \mathbb{R}^2, r > 0\}$. Can you give an example of a set A in $\mathbb{R}^2$ of size 4 that is shattered by $H_{balls}$? Give such an example or justify why you cannot find a set A of size 4 shattered by $H_{balls}$.

Solution:

We will prove that there are no four points that could be shattered by $H_{balls}$. We distinguish multiple cases:

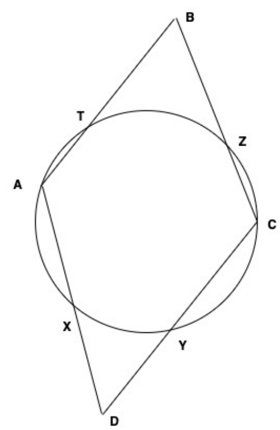
i) If all the points are collinear (suppose the order $A, B, C, D$ from left to right), then the labelling $[A, B, C, D] = [1, 0, 0, 1]$ cannot be achieved. (If 2 points are inside a circle, then the entire segment is inside the circle, and this forces $B, D$ to be labelled with 1 as well).

ii) Exactly tree points are collinear. Suppose $A, B, C$ (in this order from left to right) are the collinear ones, then again for them we cannot have the labelling $[A, B, C, D] = [1, 0, 1, 0]$ (the same as before, since $A$ and $C$ should be inside a circle, then $B$ which is inside $[A, C]$ should be inside the circle containing $A$ and $C$ as well).

iii) No three points are collinear. We split this into two cases:

iii.1) The convex hull is the triangle $\boldsymbol{\Delta} ABC$ and $D$ is inside them. In this case the labelling $[A, B, C, D] = [1, 1, 1, 0]$ cannot be obtained (Since if three points are inside a circle, then the entire triangle they form - possible degenerated to a segment - is inside the circle. This forces D to be inside the triangle as well, so its label must be 1).

iii.2) The convex hull is a convex quadrilateral $ABCD$. Since $\sphericalangle A + \sphericalangle B + \sphericalangle C + \sphericalangle D = 360°$, at least one of the pairs $(\sphericalangle A, \sphericalangle C), (\sphericalangle B, \sphericalangle D)$ have the sum $\leq 180°$. WLOG, assume this pair is $(\sphericalangle A, \sphericalangle C)$, and thus we have $\sphericalangle A + \sphericalangle C \leq 180°, \sphericalangle B + \sphericalangle D \geq 180°$. If there is a circle that manages to enclose $A$ and $C$ and let $B, D$ outside (the $[A, B, C, D] = [1, 0, 1, 0]$ case), then we can find obtain a smaller circle (the same circle if it already has the two points on its boundary) included in the previous one that contains A and C on its boundary(*). In this case, since we did not add any new points (we constructed a smaller circle that has on its boundary the points $A$ and $C$, circle that is inside the larger one - the one that labelled with 1 $A, C$ and with 0 $B, D$ -), it must be the case that $B$ and $D$ are outside the smaller circle as well, but this would force $\sphericalangle B + \sphericalangle D < 180°$ (using the formula of an exterior angle = one half of the absolute difference of the arcs it creates), a contradiction with our assumption. Thus, at least one labeling between $[1, 0, 1, 0]$ and $[0, 1, 0, 1]$ is not possible, so the four points cannot be shattered by $H_{balls}$




In conclusion, there are no four points that can be shattered by $H_{balls}$.

<u>Extra</u>: It can be easily proved that VCdim($H_{balls}$) $= 3$ by considering a set of any non-collinear 3 points. This set can indeed be shattered by $H_{balls}$

(*) <u>Proof</u>: Let $A, B$ be two points strictly inside circle $C$. Let any circle that passes through $A, B$ intersect circle $C$ in 2 points $X, Y$ (just choose a point X outside the circle and consider the circumscribed circle). Consider $\{H\} = XY \bigcap AB$ and let $T$ be one of the two contact points of the tangent lines through $H$ to circle $C$. It is a known result in geometry that the circle $HAB$ is internally tangent to $C$, so we found a smaller circle $HAB$, that passes through $A, B$ and is enclosed by circle $C$.

<u>Observation</u>: A question might arise if one point is already on the boundary of circle C, but the other is not. But in this case, we could simply enlarge the radius at an $1 + \epsilon$ rate such that the two points now are strictly inside the new circle, and the other two points are still outside. In fact, we could use any radius $r' \in [r_{initial},\ \min(|\mathrm{OB}|, |\mathrm{OD}|))$, where O is the center of the said circle. Then, we can easily apply (*).

## 3 Exercise 3

Let $X = \mathbb{R}^2$ and consider $H_\alpha$ the set of concepts defined by the area inside a right triangle $ABC$ with the two catheti $AB$ and $AC$ parallel to the axes (Ox and Oy) and with $AB/AC = \alpha$ (fixed constant $> 0$). Consider the realizability assumption. Show that the class $H_\alpha$ can be $(\epsilon, \delta)$ PAC learned by giving an algorithm A and determining an upper bound on the sample complexity $m_H(\epsilon, \delta)$ such that the definition of PAC - learnability is satisfied.

<u>Solution</u>:

Informaly, $H$ is the set of similar triangles obtained from the base triangle denoted by the points $X = (0, 0), Y = (\alpha, 0), Z = (0, 1)$ by allowing the following operations: rescale and translation.

An informal description of the algorithm $A$: For each point $P \in S$ that is labelled as 1, generate the parallel lines through it to the sides of $\triangle XYZ$. Iterating through all the points, we obtain 3 sets of lines:

$L_{XY} = \{d|\exists P \in S \text{ with label}(P) = 1, P \in d, d \parallel XY\}$
$L_{XZ} = \{d|\exists P \in S \text{ with label}(P) = 1, P \in d, d \parallel XZ\}$
$L_{YZ} = \{d|\exists P \in S \text{ with label}(P) = 1, P \in d, d \parallel YZ\}$

From $L_{XY}$ choose $d_{XY}$ that is the 'downmost', from $L_{XZ}$ choose $d_{XZ}$ that is the 'leftmost' and from $L_{YZ}$ choose $d_{YZ}$ that is 'rightmost'. (If we were to write the equation of these lines, we would notice that the elements of each set differs by a constant factor, so choose the lines with the lowest constant for $d_{XY}, d_{XZ}$ and the line with the biggest constant for $d_{YZ}$. Assume equation of a line is in the form $y = ax + b$. Here $b$ is the constant factor we are looking for). Our algorithm will return the triangle that is at the intersection of the $d_{XY}, d_{XZ}, d_{YZ}$, which is similar to the $\triangle XYZ$ (because $d_{XY}, d_{XZ}, d_{YZ}$ are parallel with the lines of the $\triangle XYZ$, it follows that this triangle is indeed in $H$). If there are no points labelled as 1, return any triangle that does not contain any points from $S$. Obviously, $L_S(h_S) = 0$.

In order to prove that $H$ is PAC-learnable, do the following: Suppose the points in the place display a distribution D (fixed), labelled by a function $h^* = \triangle ABC \in H$ (realizability setup). For the following steps, fix the $\epsilon, \delta > 0$. In our particular case, all the positive points must be inside $\triangle ABC$, while the others outside of it. An important observation of our algorithm is that it can only generate false negatives. Let $A(S) = \triangle ABC_{predicted} \subseteq \triangle ABC$ be the triangle predicted by our algorithm. Consider $R = \triangle ABC \setminus \triangle ABC_{predicted}$. We distinguish 2 cases:

i) $P[X \in ABC] < \epsilon$. In this case, $L_{D,h^*}(h_S) = P[X \in R] \leq P[X \in \triangle ABC] \leq \epsilon$, so $P_{S\sim D^m}[L_{D,h^*}(h_S) \leq \epsilon] = 1 \rightarrow P_{S\sim D^m}[L_{D,h^*}(h_S) > \epsilon] = 0 < e^{-m\epsilon},\ m \geq 0$

ii) $P[X \in ABC] \geq \epsilon$. Consider the three regions $R1, R2, R3$ defined in Figure 1 (they do exist because of $P[X \in ABC] \geq \epsilon$) such that $P[X \in R_1] = P[X \in R_2] = P[X \in R_3] = \frac{\epsilon}{3}$ (Draw parallel lines to the sides of the $h^* = \triangle ABC$ until we find the optimal positions). We have two cases to analyze:

ii.1) $\triangle ABC_{predicted} \bigcap R_i \neq \emptyset$, for all i $\in \{1, 2, 3\}$. This means that $\triangle ABC_{predicted}$ intersect all the three regions, so $L_{D,h^*}(h_S) = P[X \in R] \leq P[X \in R_1 \bigcup R_2 \bigcup R_3] \leq P[X \in R_1] + P[X \in R_2] + P[X \in R_3] = \dfrac{\epsilon}{3} + \dfrac{\epsilon}{3} + \dfrac{\epsilon}{3} = \epsilon$, so $P_{S\sim D^m}[L_{D,h^*}(h_S) \leq \epsilon] = 1 \rightarrow P_{S\sim D^m}[L_{D,h^*}(h_S) > \epsilon] = 0 < e^{-m\epsilon},\ m \geq 0$

ii.2) Let $A_i$ be the event that $\triangle ABC_{predicted} \bigcap R_i = \emptyset$, for i in $\{1, 2, 3\}$. Then, $P[L_{D,h^*}(h_S) > \epsilon] \leq P[A_1 \bigcup A_2 \bigcup A_3] \leq P[A_1] + P[A_2] + P[A_3]$ Now, $P[A_i] = P[S \bigcap R_i = \emptyset]$ (the probability of not sampling m elements from region i, which has p=$\frac{\epsilon}{3}$) $\leq (1 - \frac{\epsilon}{3})^m$. Combining these inequalities, we obtain $P[L_{D,h^*}(h_S) > \epsilon] \leq 3 * (1 - \frac{\epsilon}{3})^m < 3e^{-\frac{\epsilon m}{3}}$ $(e^x > x + 1)$. By forcing $3e^{-\frac{\epsilon m}{3}} \leq \delta \to m_H(\epsilon, \delta) = \lceil \frac{3}{\epsilon} \ln \frac{3}{\delta} \rceil$. The runtime complexity is linear in $m_H$ (up to a constant factor imposed by the dimension d = 2).

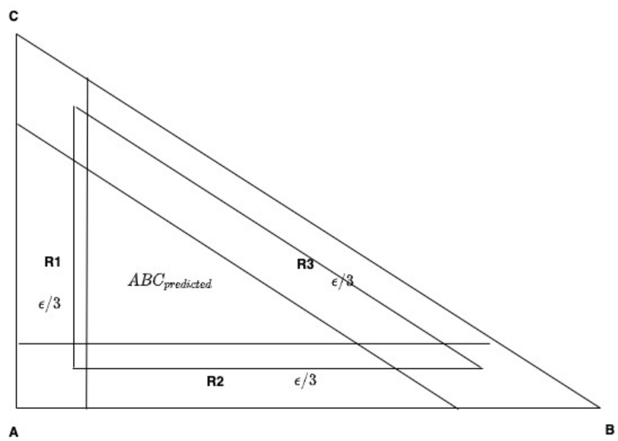



Figure 1. Spliting concept triangle into regions.

# 4 Exercise 4

Consider $H$ to be the class of all centered in origin sphere classifiers in the 3D space. A centered in origin sphere classifier in the 3D space is a classifier $h_r$ that assigns the value 1 to a point if and only if it is inside the sphere with radius $r > 0$ and center given by the origin $O(0, 0, 0)$. Consider the realizability assumption.

a. Show that the class $H$ can be $(\epsilon, \delta)$ PAC learned by giving an algorithm A and determining an upper bound on the sample complexity $m_H(\epsilon, \delta)$ such that the definition of PAC-learnability is satisfied.

b. compute VCdim$(H)$.

Solution:

An informal description of the algorithm $A$: the tightest sphere centered in origin that encapsulates all the points that are labelled as 1, or if there are no such points, a sphere centered in origin that leaves all the negative points outside of it: choose the radius equal to half the minimum distance of all the points to the origin.

Formally, let $S = \{(\mathrm{P}_1, \mathrm{label}_1), ..., (\mathrm{P}_m, \mathrm{label}_m)\}$ be the samples given to our algorithm A. Choose $r = \max(\{|\mathrm{OP}_i| | (\mathrm{P}_i, \mathrm{label}_i) \in S$ such that $\mathrm{label}_i = 1$, for $i \in \{1, ..., m\}\})$. If there are no points labelled as 1, let $r_{min} = \min(\{|\mathrm{OP}_i| | (\mathrm{P}_i, \mathrm{label}_i) \in S$ such that $\mathrm{label}_i = 0$, for $i \in \{1, ..., m\}\})$ and return $r = \frac{r_{min}}{2}$. (Because we are in the realizability conditions, we now that we cannot have $r_{min} = 0$, otherwise no sphere from $H$ can be the real labelling function).

In order to prove that $H$ is PAC-learnable, do the following: Suppose the points in the place display a distribution D (fixed), labelled by a function $h^* \in H$ (realizability setup). For the following steps, fix the $\epsilon, \delta > 0$. In our particular case, this implies that there is a radius $r_{true}$ such that the sphere centered in the origin having the radius $r_{true}$ correctly classify any possible subset of $\mathbb{R}^3$. Let $r_{predicted}$ be the radius predicted by our algorithm A.

Consider $r_{\_} = \inf(\{r | P[X \in \text{ Sphere(O, r}_{true}) \setminus \text{Sphere}(O, r)] < \epsilon, r_{true} \geq r \geq 0\})$ (the lowest radius r such that the probability of choosing a point $P \in \text{Sphere}(O, r_{true}) \setminus \text{Sphere}(O, r)$ is less than $\epsilon$).

If $r_{\_} = 0$, it follows that $P[X \in \text{ Sphere(O, r}_{true})] \leq \epsilon$, and since errors can be made only in the region $R = \text{Sphere}(O, r_{true}) \setminus \text{Sphere}(O, r_{predicted})$ (false negatives), we have $L_{D,h^*}(h_S) = P[X \in R] \leq P[X \in \text{Sphere}(O, r_{true})] \leq \epsilon$, so $P_{S \sim D^m}[L_{D,h^*}(h_S) \leq \epsilon] = 1 \rightarrow P_{S \sim D^m}[L_{D,h^*}(h_S) > \epsilon] = 0 < e^{-m\epsilon},\ m \geq 0$

If $r_{\_} > 0$, then we have $P[L_{D,h^*}(h_S) > \epsilon] = \{P[X \in \text{Sphere}(O, r_{true}) \setminus \text{Sphere(O}, r_{predicted})] > \epsilon\} \leq P[r_{predicted} \leq r_{\_}] \leq P[\mathrm{OP}_i \leq r_{\_}$, where $\mathrm{label}_i = 1, i \in \{1, ..., m\}] = \prod_1^m P[\mathrm{OP}_i \leq r_{\_}] = \prod_1^m (1 - P[r_{true} \geq \mathrm{OP}_i \geq r_{\_}]) \leq (1 - \epsilon)^m < e^{-m\epsilon}$. Letting $\delta > e^{-m\epsilon} \rightarrow m_H(\epsilon, \delta) = \lceil \frac{1}{\epsilon} \ln \frac{1}{\delta} \rceil$. The runtime complexity is linear in $m_H$ (up to a constant factor because of the dimension d=3).

b) We will prove that VCdim($H$)= 1. In order to do this, there are two steps involved:

i) Find a set consisting of a single point $S = \{P\}$ in $\mathbb{R}^3$ that is shattered by $H$. Let $P = (1, 0, 0)$. For the label 0, choose any $r_0 < 1$. Since $|OP| = 1 > r_0$ , this would imply $P$ is outside the sphere centered in origin O with radius $r_0$, so its label will be 0, as desired. Similarly, for the label 1 choose any $r_1 > 1$. But in this case, $|OP| = 1 < r_1$ and then P lies inside the sphere centered in origin O with radius $r_1$, so its label will be 1.

ii) Prove that no set of two points $S = \{A, B\}$ can be shattered by $H$. There are two cases involved:

ii.1) $|OA| = |OB|$. This means that A and B will have the same label no matter the choice of $r$, so the labeling [0, 1] cannot be achieved.

ii.2) $|OA| < |OB|$. In this case, the labelling [A,B] = [0, 1] cannot be achieved. Indeed, since A will be inside any centered sphere containing B, it must be the case that $h_r(B) = 1 \to h_r(A) = 1$, so [A,B] = [0, 1] cannot occur.

Combining the last two arguments, it follows that VCdim($H$)= 1.

# 5 Exercise 5

Let $H = \{h_\theta : \mathbb{R} \to \{0, 1\}, h_\theta(x) = 1_{[\theta,\theta+1] \bigcup [\theta+2,\infty)}, \theta \in \mathbb{R}\}$. Compute VCdim($H$).

Solution:

We will prove that VCdim$(H) = 3$.

i) For any 4 points $a_1 < a_2 < a_3 < a_4$, the labelling [1,0,1,0] cannot be achieved by a classifier in $H$.
Proof: Suppose that exists $h_\theta \in H$ that can generate the previous labelling. Since $h(a_1) = 1$, it must be the case that $a_1$ lies either inside $[\theta, \theta + 1]$ or $[\theta + 2, \infty)$. The last case fails as that would imply all the remaining points will be labelled as 1, since all of them are greater than $a_1$. So we will focus on the first case, namely $a_1 \in [\theta, \theta + 1]$. Since the only region that could label a point 0 on the right of $a_1$ is the region $(\theta + 1, \theta + 2)$, we need that $a_2 \in (\theta + 1, \theta + 2)$. But then, since $h_\theta(a_3) = 1$, and the only region on the right of $a_2$ that could give label 1 is $[\theta + 2, \infty)$, it must be the case that $a_3 \in [\theta + 2, \infty)$. Finally, since all the points on the right of $a_3$ will be labelled as 1, we cannot have $h_\theta(a_4) = 0$, a contradiction. In conclusion, there are no set of four points that could be shattered by $H$.

ii) Let $S = \{a_0 = 0, a_1 = 0.6, a_2 = 1.3\}$. We distinguish eight cases:

1) [label($a_0$), label($a_1$), label($a_2$)] $= [0, 0, 0]$. Choose $\theta = 1.4 \to 1_{[1.4,2,4] \bigcup [3.4,\infty)}$

2) [label($a_0$), label($a_1$), label($a_2$)] $= [0, 0, 1]$. Choose $\theta = 1.2 \to 1_{[1.2,2.2] \bigcup [3.2,\infty)}$

3) [label($a_0$), label($a_1$), label($a_2$)] $= [0, 1, 0]$. Choose $\theta = 0.2 \to 1_{[0.2,1.2] \bigcup [2.2,\infty)}$

4) [label($a_0$), label($a_1$), label($a_2$)] $= [0, 1, 1]$. Choose $\theta = 0.5 \to 1_{[0.5,1.5] \bigcup [2.5,\infty)}$

5) [label($a_0$), label($a_1$), label($a_2$)] $= [1, 0, 0]$. Choose $\theta = -0.5 \to 1_{[-0.5,0.5] \bigcup [1.5,\infty)}$

6) [label($a_0$), label($a_1$), label($a_2$)] $= [1, 0, 1]$. Choose $\theta = -0.8 \to 1_{[-0.8,0.2] \bigcup [1.2,\infty)}$

7) [label($a_0$), label($a_1$), label($a_2$)] $= [1, 1, 0]$. Choose $\theta = -0.1 \to 1_{[-0.1,0.9] \bigcup [1.9,\infty)}$

8) [label($a_0$), label($a_1$), label($a_2$)] $= [1, 1, 1]$. Choose $\theta = -2.1 \to 1_{[-2.1,-1.1] \bigcup [-0.1,\infty)}$

According to the last eight cases, it follows that the set $S$ is shattered by $H$, and since no set of four points can be shattered by $H$, it implies that VCdim$(H) = 3$.

## 6 Exercise 6

Let X be an instance and consider $H \subseteq \{0,1\}^X$ a hypothesis space with finite VC dimension. For each $x \in X$, we consider the function $z_x : H \to \{0,1\}$ such that $z_x(h) = h(x)$ for each $h \in H$. Let $Z = \{z_x : H \to \{0,1\}, x \in X\}$. Prove that $VCdim(\mathrm{Z}) < 2^{VCdim(H)+1}$.

Solution:

We will approach this problem by assuming the contrary and proving that this gives us a contradiction. Let's assume that VCdim$(Z) \geq 2^{\mathrm{VCdim}(H)+1}$, and let d = VCdim$(H)$. In this case, there exists a subset $S \subseteq H$ with $|S| =$ VCdim$(Z) \geq 2^{d+1}$ that is shattered by Z. From that subset, extract a smaller subset $A = \{h_1, ..., h_{2^{d+1}}\} \subseteq S$, such that $|A| = 2^{d+1}$ (this can be always done, as shattering a set implies that every subset can be shattered as well). This implies that for each possible labelling $L \in \{0,1\}^{2^{d+1}}$, $\exists z_{x_L} \in Z$ such that $z_{x_L}(A) = \{z_{x_L}(h_1), ..., z_{x_L}(h_{2^{d+1}})\} = \{h_1(x_L), ..., h_{2^{d+1}}(x_L)\} = L$. Of course, the mapping $L \to x_L$ is injective. Let $M \in \{0,1\}^{d+1 \times 2^{d+1}}$ be the matrix that has as the $i^{th}$ column the binary representation of $i$, from top (least significant bit) to bottom (most significant bit) (0-indexes). Now, since each row of this matrix can be viewed as a possible labelling in $\{0,1\}^{2^{d+1}}$, just get the corresponding $L = r_i \to x_{r_i}$(A problem that might occur is that if there exists two rows in $M$ that are equals, but this cannot be the case. If this would hold, say that $r_i = r_j$, then this would imply that the numbers generated per column would always have bit i and bit j equal, which cannot hold since each column denote the binary representation of its position in matrix $M$).

$$\begin{bmatrix} h_1(x_{r_0}) & h_2(x_{r_0}) & \dots & h_{2^{d+1}}(x_{r_0}) \\ h_1(x_{r_1}) & h_2(x_{r_1}) & \dots & h_{2^{d+1}}(x_{r_1}) \\ \vdots & \vdots & \ddots & \vdots \\ h_1(x_{r_d}) & h_2(x_{r_d}) & \dots & h_{2^{d+1}}(x_{r_d}) \end{bmatrix} = \begin{bmatrix} \mathrm{binary}(0) & \mathrm{binary}(1) & \dots & \mathrm{binary}(2^{d+1}-1) \end{bmatrix}$$

But this last equation proves that the set $S = \{x_{r_0}, ..., x_{r_d}\}$ is shattered by $H$, and since $|S| = d + 1$, it implies that VCdim$(H) \geq d + 1$, a contradiction with our hypothesis that assumed VCdim$(H) = d$. The desired conclusion follows.