Advanced Machine Learning

Assignment – 1

Tomeh Lara

April 2022

Question – 1

Give an example of a finite hypothesis class $\mathcal{H}$ with *VCdim* $(\mathcal{H})$ *= 2022*. Justify your choice.

From lecture 6 we know that in order to show that the *VC-dimension* of a hypothesis class $\mathcal{H}$ is d, we need to show that:

1- There exists a set C of size d that is shattered by $\mathcal{H}$. (*VCdim* $(\mathcal{H}) \geq d$).
2- Every set C of size d + 1 is not shattered by $\mathcal{H}$. (*VCdim* $(\mathcal{H}) < d+1$)

We have a set $C = \{c1, c2, \ldots, c\eta\}$ the orthonormal basis of $\mathbb{R}^n$, the finite hypothesis class $\mathcal{H}$ is: $\mathcal{H} = \{h_{w,o}: \mathbb{R}^n \rightarrow \{-1,1\}\ ,\ h_{w,o} = sign\ (\sum_{i=1}^{n} w_i x_i) \mid w_i = \begin{cases} 1, & if\ c_i \in D \\ -1, & if\ c_i \notin D \end{cases}$ *for all* subsets D of C}. The evidence consists of two steps:

Step-1. Prove that $VCdim\ (\mathcal{H}) \geq 2022$, We have set C of n points in $\mathbb{R}^n$ that is shattered by $\mathcal{H}$ and we have to prove that for every subset D of C there is a function h_D that labels +1 all elements in D and -1 all elements not in D. $\mathcal{H}$ contains one hypothesis for every subset D in C:

$$h_D = h_{w,o} = sign\ (\sum_{i=1}^{2022} w_i x_i) \text{ where } w_i = \begin{cases} 1\ , & if\ c_i \in D \\ -1, & if\ c_i \notin D \end{cases}$$

We have $h_D = sign\ (\langle w, e_i \rangle) = w_i$, consequently it assigns +1 for all elements in D and -1 for all elements not in D, so now we proved that $VCdim\ (\mathcal{H}) \geq n$ where $n = 2022$ in this case.

Step-2. Prove that $VCdim\ (\mathcal{H}) < n + 1$. For a finite hypothesis class $\mathcal{H}$ we have the upper bound: *VCdim* $(\mathcal{H}) \leq \lceil log_2(|\mathcal{H}|) \rceil$. Now in this exercise we have one hypothesis for each subset of C, thus $|\mathcal{H}| = 2^{|C|} = 2^n$. As a result: $VCdim\ (\mathcal{H}) \leq \lceil log_2(2^n) \rceil = n$.
We conclude the proof by combining the previous two steps:

$$h_D = h_{w,o} = (\sum_{i=1}^{2022} w_i x_i) \text{ where } w_i = \begin{cases} 1, & if\ c_i \in D \\ -1, & if\ c_i \notin D \end{cases}$$

Thus, VC Dim of 2022 half spaces is equal to 2022.

$$VC\ \dim(\mathcal{H}) = n \ \Rightarrow\ VC\ \dim(\mathcal{H}) = 2022\ .$$

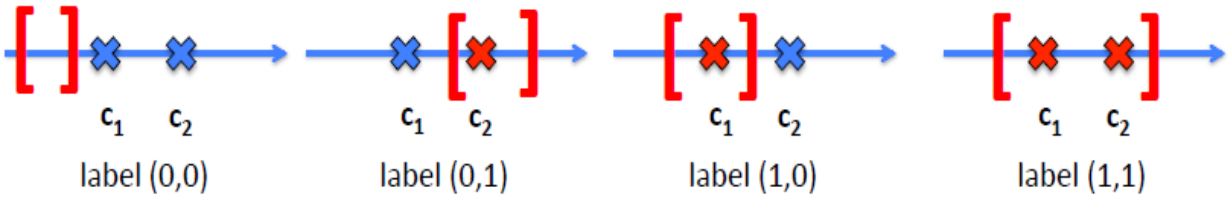
Question – 2

What is the maximum value of the natural even number $n, n = 2m$, such that there exists a hypothesis class $\mathcal{H}$ with n elements that shatters a set C of $m = \frac{n}{2}$ points? Give an example of such an $\mathcal{H}$ and $\mathcal{C}$. Justify your answer.

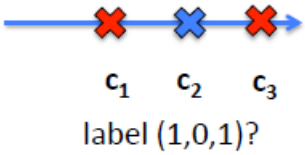
$\mathcal{H}_{intervals} = \{[a,b] |\ a \leq b\ , a, b \in \mathbf{R}\}$ can also view as:

$\mathcal{H}_{intervals} = \{h_{a,b}: \mathbf{R} \rightarrow \{0,1\}\ , h_{a,b} = 1_{[a,b]}, a \leq b\ \ , a, b \in \mathbf{R}\}$

$\mathcal{H}_{intervals}$ shatters any set A of two different points in **R**.



$\mathcal{H}_{intervals}$ cannot shatter any set A of three different points in **R**.



Therefore, $VC\ \dim(\mathcal{H}_{intervals}) = 2$

$|\mathcal{H}_{intervals}| = 2^{|C|} = 2^2 = 4$. Thus, $|\mathcal{H}_{intervals}| = 2 \times VC\ \dim(\mathcal{H}_{intervals})$

We will use the growth function measures the maximal "effective" size of $\mathcal{H}$ on a set of m examples. Formally:

Let $\mathcal{H}$ be a hypothesis class. Then the *growth function* of $\mathcal{H}$, denoted $\tau_{\mathcal{H}}: \mathbb{N} \rightarrow \mathbb{N}$,is defined as

$$\tau_{\mathcal{H}}(m) = \max_{C \subset \mathcal{X}: |C| = m} |\mathcal{H}_C|$$

In words, $\tau_{\mathcal{H}}(m)$ is the number of different functions from a set C of size m to $\{0,1\}$ that can be obtained by restricting $\mathcal{H}$ to C.

Obviously, if $VC\ \dim(\mathcal{H}) = d$,then for any $m \leq d$ we have $\tau_{\mathcal{H}}(m) = 2^m$. In such cases, $\mathcal{H}$ induces all possible functions from C to $\{0,1\}$. The Sauer's Lemma shows that when m becomes larger than the VC-dimension, the growth function increases polynomially rather than exponentially with m.

Let $\mathcal{H}$ be *a hypothesis class with* VCdim $(\mathcal{H}) \leq d\ < \infty$. Then, for all m, $\tau_{\mathcal{H}}(m)\ \leq \sum_{i=0}^{d} \binom{m}{i}$.

In particular, if $m\ > d + 1$ then $\tau_{\mathcal{H}}(m) \leq (em/d)^d$.

Let m, d *be two positive integers such that* $d \leq m - 2$. *Then,*

$$\sum_{k=0}^{d} \binom{m}{k} \leq \left(\frac{e\ m}{d}\right)^d$$

We prove the claim by induction. For $d = 1$ the left-hand side equals $1 + m$ while the right-hand side equals em; hence the claim is true. Assume that the claim holds for d and let us prove it for d +1. By the induction assumption we have

$$\sum_{k=0}^{d+1}\binom{m}{k} \le \left(\frac{e\,m}{d}\right)^d + \binom{m}{d+1}$$

$$= \left(\frac{e\,m}{d}\right)^d \left(1 + \left(\frac{d}{e\,m}\right)^d \frac{m(m-1)(m-2)\ldots(m-d)}{(d+1)d!}\right)$$

$$\le \left(\frac{e\,m}{d}\right)^d \left(1 + \left(\frac{d}{e}\right)^d \frac{(m-d)}{(d+1)d!}\right)$$

Using Stirling's approximation, we further have that

$$\le \left(\frac{e\,m}{d}\right)^d \left(1 + \left(\frac{d}{e}\right)^d \frac{(m-d)}{(d+1)\sqrt{2\pi d}(d/e)^d}\right)$$

$$= \left(\frac{e\,m}{d}\right)^d \left(1 + \frac{m-d}{\sqrt{2\pi d}(d+1)}\right)$$

$$= \left(\frac{e\,m}{d}\right)^d \cdot \left(\frac{d+1(m-d)/\sqrt{2\pi d}}{d+1}\right)$$

$$\le \left(\frac{e\,m}{d}\right)^d \cdot \left(\frac{d+1+(m-d)/2}{d+1}\right)$$

$$= \left(\frac{e\,m}{d}\right)^d \cdot \left(\frac{d/2+1+m/2}{d+1}\right)$$

$$\le \left(\frac{e\,m}{d}\right)^d \cdot \frac{m}{d+1}$$

where in the last inequality we used the assumption that $d \le m - 2$. On the other hand,

$$\left(\frac{e\,m}{d+1}\right)^{d+1} = \left(\frac{em}{d}\right)^d \cdot \frac{em}{d+1} \cdot \left(\frac{d}{d+1}\right)^d$$

$$= \left(\frac{em}{d}\right)^d \cdot \frac{em}{d+1} \cdot \frac{1}{(1+1/d)^d}$$

$$\ge \left(\frac{em}{d}\right)^d \cdot \frac{em}{d+1} \cdot \frac{1}{e}$$

$$= \left(\frac{em}{d}\right)^d \cdot \frac{m}{d+1}$$

which proves our inductive argument.

Question – 3

Let $\mathcal{X} = \mathbb{R}^2$ and consider $\mathcal{H}$ the set of axes aligned rectangles with the center in origin *O (0, 0)*. Compute the $VC\dim(\mathcal{H})$.

A set of axes aligned rectangles with the center in origin is known as:

$$\mathcal{H}_{rec}^2 = \left\{ h_{(a_1,b_1,a_2,b_2)} \middle|\ a_i = -b_i, i = 1,2 \right.$$

$$h_{(a_1,b_1,a_2,b_2)}(\underline{x}) = \begin{cases} 1, a_i \le x_i \le b_i, \forall_i = 1,2 \\ 0, \ otherwise \end{cases}$$

We have two steps:

Step-1: Prove $VC\dim(\mathcal{H}) \ge 1$.

Let $x \in \mathbb{R}^2$, $h(x) = 1$ $if - a_1 \le x \le a_1, and - a_2 \le x \le a_2$

Step-2: Prove $VC\dim(\mathcal{H}) < 2$

Let $x_1, x_2 \in \mathbb{R}^2, O \le |x_1| \le |x_2|$

In the hypothesis $\mathcal{H}$ we cannot classify the points those labelled (0,1) using any classifier without misclassification x, therefore $VC\dim(\mathcal{H}) = 1$

Step-1. Show that the VCdim $\ge 2d$. Consider a set of 2d points where each point only has one of the d dimensions set to either 1 or −1 and 0 for all other dimensions. It is easy to see that any subset of these points can be shattered by an axis-aligned rectangle. Hence the VCdim is at least $2d$.

Let $C = \{c_1, c_2, c_3, \dots, c_{2d-1}, c_{2d}\}$ where

$c_1 = (1,0,0,\dots,0) = e_1$

$c_2 = (0,1,0,\dots,0) = e_2$

….

$c_d = (0,0,0,\dots,1) = e_d$ $\qquad$ $c_i = e_i = -c_{i+d}$

$c_{d+1} = (-1,0,0,\dots,0) = -e_1$ $\qquad$ $\forall_i = \overline{1,d}$

$c_{d+2} = (0,-1,0,\dots,0) = -e_2$

….

$c_{2d} = (0,0,0,\dots,-1) = -e_d$

For d = 2, we have in 2 dimensions: $c_1 = (1,0)$, $c_2 = (0,1)$, $c_3 = (-1,0)$, $c_4 = (0,-1)$,

and for each labeling $(y_1, y_2, \dots, y_{2d})$ of the points $(c_1, c_2, \dots, c_{2d})$, (there are $2^{2\mathrm{d}}$ possible labelings), there exists a function h in $\mathcal{H}_{\mathrm{rec}}^{\mathrm{d}}$ such that $\mathrm{h}(c_i) = y_i \; \forall i = \overline{1, 2d}$.

Consider a labeling $(y_1, y_2, \dots, y_{2d}) \in \{0,1\}^{2\mathrm{d}}$. Each point c_i has all components = 0, apart from component $i \; if \; i \in \{1, \dots, \mathrm{d}\}$ or $i - d \; if \; i \in \{\mathrm{d} + 1, \dots, 2\mathrm{d}\}$.

The choice of the interval $[a_i, b_i]$ is influenced by the labels y_i and y_{i+d} of the points c_i and c_{i+d}. As all other points $c_1, c_2, \dots, c_{i-1}, c_{i+1}, \dots, c_{i+d-1}, c_{i+d+1}, \dots, c_{2d}$ have 0 on the *i-th* component, we have that $[a_i, b_i]$ should contain 0, otherwise each point will be labeled with 0.

So $[a_i, b_i]$ depends on y_i and y_{i+d}, and $[a_i, b_i]$ decides basically the label of points c_i and c_{i+d}:
$c_i = (0, \dots, 0, 1, 0, \dots, 0) \qquad c_{i+d} = (0, \dots, 0, -1, 0, \dots, 0)$

The possible cases:

- $y_i = 0, \; y_{i+d} = 0, then \; [a_i, \; b_i] \cap \{-1,1\} = \phi$
  $[a_i, b_i]$ should not contain points -1 and 1.
  In this case, take $a_i = -0.5, \; b_i = 0.5$
- $y_i = 0, \; y_{i+d} = 1, then \; [a_i, \; b_i] \cap \{-1,1\} = \{-1\}$
  $[a_i, b_i]$ should contain only point -1 such that c_{i+d} will get label 1.
  In this case, take $a_i = -2, \; b_i = 0.5$
- $y_i = 1, \; y_{i+d} = 0, then \; [a_i, \; b_i] \cap \{-1,1\} = \{1\}$
  $[a_i, b_i]$ should contain only point +1 such that c_i will get label 1.
  In this case, take $a_i = -0.5, \; b_i = 2$
- $y_i = 1, \; y_{i+d} = 1, then \; [a_i, \; b_i] \cap \{-1,1\} = \{-1,1\}$
  $[a_i, b_i]$ should contain both points {-1,1} such that c_i and c_{i+d} will get label 1.
  In this case, take $a_i = -2, \; b_i = 2$

In all cases, we have that $h_{(a_1, \; b_1, \; a_2, \; b_2, \dots, a_d, \; b_d)}(c_i) = y_i, \; \forall_i = \overline{1,2d}$, where each interval $[a_i, \; b_i]$ was determined based on $y_i \; and \; y_{i+d}$, $i = \overline{1, d}$, therefore $VC \dim(\mathcal{H}_{rec}^{d}) \geq 2d$.

Step-2. Show that the VC–dim < $2d$ + 1. Consider a set of 2d + 1 points. Consider finding the minimum and maximum of value in each dimension for these set of points and then building a Rd rectangle with these bounds. Since there are 2d + 1 points, at least one point must lie inside this rectangle. If we label this interior point as negative then there is no rectangle that can separate this labeling. This proves that VC–dim < 2d + 1.

Let C be a set of size $2d + 1$ points and because we have $2d + 1$ points in C and there are only d dimensions, there will exist a point $x \in C$ such that, for each dimension $i = \overline{1, d}$ there will be 2 points x' and $x'' \in$ C such that $x'_i \leq x \leq x''_i$ (the point x_i is "inside" the convex hull determined by all other points in dimension i). So, the label for which x has value 0 and all other 2d points get label 1 cannot be realized by any function $h \in \mathcal{H}_{rec}^{d}$ (because x is inside the rectangle) that contain all other points. Combining the previous two steps we get that the

VCdim = $2d$.

$\mathcal{H}_{rec}^{d} = \{h_{(a_1,b_1,a_2,b_2,\ldots,a_d,b_d)} |\ a_i \leq b_i, i = \overline{1,d}\}$

$$h_{(a_1,b_1,a_2,b_2,\ldots,a_d,b_d)}(\underline{x}) = \begin{cases} 1, & a_i \leq x^i \leq b_i \quad \forall_i = \overline{1,d} \\ 0, & otherwise \end{cases}$$

$\mathcal{H} = \mathcal{H}_{rec}^{2}$, the set of axes aligned rectangles in the $\mathbb{R}^2$.

$\mathcal{H}_{rec}^{2} = \{[a,b] \times [c,d] |\ a \leq b, c \leq d, a, b, c, d \in \boldsymbol{R}\}$

$\mathcal{H}_{rec}^{2} = \{h_{a,b,c,d} : \boldsymbol{R}^2 \to \{0,1\}, h_{a,b,c,d} = 1_{[a,b]\times[c,d]}, a \leq b, c \leq d, a, b, c, d \in \boldsymbol{R}\}$

$\mathcal{H}_{rec}^{2}$ shatters the set A of 4 points arranged as a diamond, that's why $VC\dim(\mathcal{H}_{rec}^{2}) \geq 4$.

$\mathcal{H}_{rec}^{2}$ does not shatter any set A of five points in $\mathbb{R}^2$, therefore, $VC\dim(\mathcal{H}_{rec}^{2}) = 4$.

Knowing that the VC dimension of rectangles is the cardinality of the maximum set of points that can be shattered by a rectangle and the VC dimension of rectangles is 4 because there exists a set of 4 points that can be shattered by a rectangle and any set of 5 points cannot be shattered by a rectangle. So, while it's true that a rectangle cannot shatter a set of four collinear points with alternate positive and negative, the VC-dimension is still 4 because there exists one configuration of 4 points which can be shattered.

Question – 4

Let $\mathcal{X} = \mathbb{R}^2$ and consider $\mathcal{H}_\alpha$ the set of concepts defined by the area inside a right triangle ABC with two catheti AB and AC parallel to the axes (Ox and Oy), and with the ratio AB/AC = α (fixed constant > 0). Consider the realizability assumption. Show that the class $\mathcal{H}_\alpha$ is (ϵ, δ)-PAC learnable 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.

$\mathcal{H}$ is a set of triangles from the base one expressed by the points $\mathcal{X} = (0,0)$, $\mathcal{Y} = (a, 0)$, $\mathcal{Z} = (0,1)$ through rescale and translation operations. The algorithm A claimed that 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:

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

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

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

We pick the downmost d_{XY} from L_{XY} , the leftmost d_{XZ} from L_{XZ} , and the rightmost d_{YZ} from L_{YZ}. We have to choose the lines with the lowest constant for d_{XY} and d_{XZ} and the line with the biggest constant for d_{YZ} because the elements of each set differs by a constant factor. The equation form is $\mathcal{Y} = ax + b$ and the algorithm will return the triangle that is at the intersection of the picked points which is similar to $\triangle XYZ$, so if there are no points labelled as 1, it will give any triangle that does not contain any point from S, thus, $L_S(h_S) = 0$. Assume that the points in the place display a distribution D labelled by a function $h^* = \triangle ABC\ \in H$, and $\epsilon, \delta > 0$, to prove that H is PAC-learnable. In this situation all the positive points have to be inside $\triangle ABC$ and the others outside of it, therefore our algorithm can only generate false negatives. Assume $A(S) = \triangle ABC_{predicted}\ \subseteq\ \triangle ABC$, and $R = \triangle ABC \backslash \triangle ABC_{predicted}$. We discriminate the following two occurrences:

1. $P[X \in ABC]\ < \epsilon$. In this situation, $L_{D,h^*}(h_S) = P[X \in R] \leq P[X \in \triangle ABC]\ \leq\ \epsilon$, so $P_{S\sim D^m}\left[L_{D,h^*}(h_S)\ \leq\ \epsilon\right] = 1\ \rightarrow\ P_{S\sim D^m}\left[L_{D,h^*}(h_S)\ >\ \epsilon\right] = 0\ <\ e^{-m\epsilon}, m \geq 0$
2. $P[X \in ABC]\ \geq \epsilon\ such\ that\ P[X \in R_1] = P[X\ \in R_2] =\ P[X\ \in R_3] = \frac{\epsilon}{3}$. Take into consideration the three regions $R_1, R_2, R_3\ in\ figure\ 1$, Now if we draw parallel lines to the sides of the $h^* = \triangle ABC$ until we get the optimal positions, we got the following cases:
   Case-1: $\triangle ABC_{predicted}\ \cap\ R_i\ \neq \emptyset$, for all $i\ \in \{1,2,3\}$. This means that $\triangle ABC_{predicted}$ intersect R_1, R_2, R_3 , therefore,
   $L_{D,h^*}(h_S) =\ P[X \in R]\ \leq\ P[X \in R_1 \cup R_2 \cup R_3]\ \leq P[X \in R_1] +\ P[X \in R_2] +\ P[X \in R_3] = 3\left(\frac{\epsilon}{3}\right) =\ \epsilon$, therefore
   $P_{S\sim D^m}\left[\ L_{D,h^*}(h_S)\ \leq\ \epsilon\right] = 1\ \rightarrow\ P_{S\sim D^m}\left[L_{D,h^*}(h_S) > \epsilon\right] = 0\ <\ e^{-m\epsilon}, m \geq 0.$

Case-2: Assume A_i be the event that $\triangle ABC_{predicted} \cap R_i \neq \emptyset$, for $i\ in\ \{1,2,3\}$. Then, $P[L_{D,h^*}(h_s) > \epsilon] \leq P[A_1 \cup A_2 \cup A_3] \leq P[A_1] + P[A_2] + P[A_3]$.

$P[A_i] = P[S \cap R_i = \emptyset]$, the probability of not sampling m elements from region i,

Which has $p = \frac{\epsilon}{3} \leq (1 - \frac{\epsilon}{3})^m$.

Merging these two dissimilarities, we gain:

$$P[L_{D,h^*}(h_s) > \epsilon] \leq 3 * \left(1 - \frac{\epsilon}{3}\right)^m < 3e^{-\frac{\epsilon m}{3}} (e^x > x + 1).$$

If $3e^{-\frac{\epsilon m}{3}} \leq \delta \rightarrow m_H(\varepsilon, \delta) = \left\lceil \frac{3}{\epsilon}\ in \frac{3}{\delta} \right\rceil$.

The runtime complexity is linear in m_H ,up to a constant factor imposed by the dimension $d = 2$.

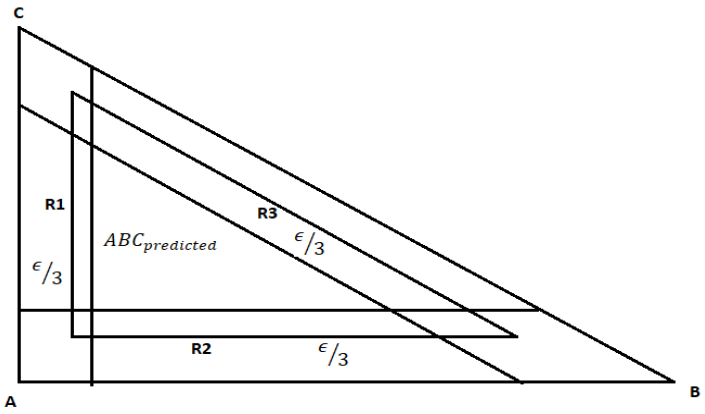


Figure 1 Splitting triangle into regions

Question – 5

Consider $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2 \cup \mathcal{H}_3$, where:

$\mathcal{H}_1 = \{h_{\theta_1} : \mathbb{R} \longrightarrow \{0,1\} \mid h_{\theta_1}(x) = 1_{[x \geq \theta_1]}(x) = 1_{[\theta_1, +\infty)}(x), \theta_1 \in \mathbb{R}\}$,

$\mathcal{H}_2 = \{h_{\theta_2} : \mathbb{R} \longrightarrow \{0,1\} \mid h_{\theta_2}(x) = 1_{[x < \theta_2]}(x) = 1_{(-\infty, \theta_2)}(x), \theta_2 \in \mathbb{R}\}$,

$\mathcal{H}_3 = \{h_{\theta_1,\theta_2} : \mathbb{R} \longrightarrow \{0,1\} \mid h_{\theta_1,\theta_2}(x) = 1_{[\theta_1 \leq x \leq \theta_2]}(x) = 1_{[\theta_1, \theta_2]}(x), \theta_1, \theta_2 \in \mathbb{R}\}$.

Consider the realizability assumption.

a) Compute $VC\dim(\mathcal{H})$.

b) Show that $\mathcal{H}$ is PAC-learnable.

c) Give an algorithm A and determine an upper bound on the sample complexity $m_{\mathcal{H}}(\epsilon, \delta)$ such that the definition of PAC-learnability is satisfied.

a) Compute $VC\dim(\mathcal{H})$

Let $\mathcal{H}_1, \mathcal{H}_2, \mathcal{H}_3$ be hypothesis classes over some fixed domain set $\mathcal{X}$. Let $d = max_i\, VCdim(\mathcal{H}_i)$ and assume for simplicity that $d \geq 3$. VCdim $(\cup_{i=1}^{r} \mathcal{H}_i) \leq 4d \log(2d) + 2\log(r)$. Take a set of *k* examples and assume that they are shattered by the union class. Therefore, the union class can produce all 2^k possible labelings on these examples. Use Sauer's lemma to show that the union class cannot produce more than $r\, k^d$ labelings. Therefore, $2^k \; r\, k^d$.

Sauer's lemma: Let $\mathcal{H}$ be *a hypothesis class with* VCdim $(\mathcal{H}) \leq d < \infty$. Then, for all m,

$\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$. In particular, if $m > d + 1$ then $\tau_{\mathcal{H}}(m) \leq (em/d)^d$

Proof of Sauer's Lemma: For any $C = \{c_1, \ldots, c_m\} we\ have$

$$\forall \mathcal{H}, |\mathcal{H}_c| \leq |\{B \subseteq C : \mathcal{H}\ shatters\ B\}| .$$

$VC\dim(\mathcal{H}) \leq d$ then no set whose size is larger than *d* is shattered by $\mathcal{H}$ and therefore

$$|\{B \subseteq C : \mathcal{H}\ shatters\ B\}| \leq \sum_{i=0}^{d} \binom{m}{i}$$

When $m > d + 1$ the right-hand side of the preceding is at most $(em/d)^d$. Now fix $\mathcal{H}$ and $C = \{c_1, \ldots, c_m\}$. Denote $C' = \{c_2, \ldots, c_m\}$, and in addition, define the following two sets:

$$Y_0 = \{(y_2, \ldots, y_m) : (0, y_2, \ldots, y_m) \in \mathcal{H}_c \vee (1, y_2, \ldots, y_m) \in \mathcal{H}_c\},$$

And

$$Y_1 = \{(y_2, \ldots, y_m) : (0, y_2, \ldots, y_m) \in \mathcal{H}_c \wedge (1, y_2, \ldots, y_m) \in \mathcal{H}_c\}.$$

It is easy to verify that $|\mathcal{H}_c| = |y_0| + |y_1|$.

Additionally, since $y_0 = \mathcal{H}_c$ using the induction assumption (applied on $\mathcal{H}$ and C') we have that $|y_0| = |\mathcal{H}_{c'}| \leq |\{B \subseteq C' : \mathcal{H}\ shatters\ B\}| = |\{B \subseteq C : c_1 \notin B \wedge \mathcal{H}\ shatters\ B\}|$.

Define $\mathcal{H}' \subseteq \mathcal{H}$:

$\mathcal{H}' = \{h \in \mathcal{H} : \exists h' \in \mathcal{H}\ s.t. \big(1 - h'(c_1), h'(c_2), \ldots, h'(c_m)\big) = \big(h(c_1), h(c_2), \ldots, h(c_m)\big)\}$,

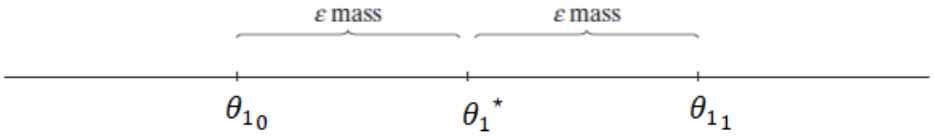
Namely, $\mathcal{H}'$ contains pairs of hypotheses that agree on C' and differ on c_1, Using this definition, it is clear that if $\mathcal{H}'$ shatters a set $B \subseteq C'$ then it also shatters the set $B \cup \{c_1\}$ and vice versa. Combining this with the fact that $Y_1 = \mathcal{H}'_{c'}$ and using the inductive assumption we obtain that

$|Y_1| = |\mathcal{H}'_{c'}| \leq |\{B \subseteq C' : \mathcal{H}' shatters\ B\}| = |\{B \subseteq C' : \mathcal{H}' shatters\ B \cup \{c_1\}\}|$
$= |\{B \subseteq C : c_1 \in B \wedge \mathcal{H}' shatters\ B\}| \leq |\{B \subseteq C : c_1 \in B \wedge \mathcal{H}\ shatters\ B\}|$
So, $|\mathcal{H}_c| = |y_0| + |y_1|$
$\leq |\{B \subseteq C : c_1 \notin B \wedge \mathcal{H}\ shatters\ B\}| + |\{B \subseteq C : c_1 \in B \wedge \mathcal{H}\ shatters\ B\}|$
$= |\{B \subseteq C : \mathcal{H}\ shatters\ B\}|$.
Now using Lemma: If $a \geq 1$ and $b > 0$. Then, $x \geq 4a\ log(2a) + 2b \implies x \geq a\ log(x) + b$.
Prove that for $r = 2$ it holds that $VC\ \dim(\mathcal{H}_1 \cup \mathcal{H}_2 \cup \mathcal{H}_3) \leq 2d + 1$. Thus, a sequence of 2k+1 points on a line cannot be shattered, if successive points are labeled with alternate labels, starting with a positive label. Thus, VC dimension of the class of union of k intervals on the real line is 2k.

b) Show that $\mathcal{H}$ is PAC-learnable.

$\mathcal{H}_1 = \{h_{\theta_1} : \mathbb{R} \longrightarrow \{0,1\} \mid h_{\theta_1}(x) = 1_{[x \geq \theta_1]}(x) = 1_{[\theta_1, +\infty)}(x), \theta_1 \in \mathbb{R}\}$
Means $h_{\theta_1}(x) = 1\ if\ x \geq \theta_1$, and 0 otherwise.
$\mathcal{H}_2 = \{h_{\theta_2} : \mathbb{R} \longrightarrow \{0,1\} \mid h_{\theta_2}(x) = 1_{[x < \theta_2]}(x) = 1_{(-\infty, \theta_2)}(x), \theta_2 \in \mathbb{R}\}$
Means $h_{\theta_2}(x) = 1\ if\ x < \theta_2$, and 0 otherwise.
$\mathcal{H}_3 = \{h_{\theta_1,\theta_2} : \mathbb{R} \longrightarrow \{0,1\} \mid h_{\theta_1,\theta_2}(x) = 1_{[\theta_1 \leq x \leq \theta_2]}(x) = 1_{[\theta_1, \theta_2]}(x), \theta_1, \theta_2 \in \mathbb{R}\}$
Means $h_{\theta_1,\theta_2}(x) = 1\ if\ \theta_1 \leq x \leq \theta_2$, and 0 otherwise.
It is obvious that $\mathcal{H}$ is of infinite size. Nevertheless, the following lemma shows that $\mathcal{H}$ is learnable in the PAC model using the ERM algorithm.
Let $\mathcal{H}$ be the class of thresholds as defined earlier. Then, $\mathcal{H}$ is PAC learnable, using the ERM rule, with sample complexity of $m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \log\left(2/\delta\right)/\epsilon \right\rceil$.
Let ${\theta_1}^\star$ be a threshold such that the hypothesis $h^\star(x) = 1_{[x \geq {\theta_1}^\star]}$ achieves $L_D(h^\star) = 0$. Let D_x be the marginal distribution over the domain $\mathcal{X}$ and let $\theta_{1_0} < {\theta_1}^\star < \theta_{1_1}$ be such that:

$$\mathbb{P}_{x \sim D_x}\left[x \in (\theta_{1_0}, {\theta_1}^\star)\right] = \mathbb{P}_{x \sim D_x}\left[x \in ({\theta_1}^\star, \theta_{1_1})\right] = \epsilon$$



(If $D_x(-\infty, {\theta_1}^\star) \leq \epsilon$ we set $\theta_{1_0} = -\infty$ and similarly for θ_{1_1}). Given a training set S, let $b_0 = \max\{x: (x,1) \in S\}$ and $b_1 = \min\{x: (x,0) \in S\}$ (if no example in S is positive we set $b_0 = -\infty$ and if no example in S is negative we set $b_1 = \infty$). Let b_s be a threshold corresponding to an ERM hypothesis, h_s, which implies that $b_s \in (b_0, b_1)$. Therefore, a sufficient condition for $L_D(h_s) \leq \epsilon$ is that both $b_0 \geq \theta_{1_0}$ and $b_1 \leq \theta_{1_1}$. In other words,

$$\mathbb{P}_{S \sim D^m}[L_D(h_s) > \epsilon] \leq \mathbb{P}_{S \sim D^m}\left[b_0 < \theta_{1_0} \vee b_1 > \theta_{1_1}\right],$$

and using the union bound we can bound the preceding by

$$\mathbb{P}_{S\sim D^m}[L_D(h_s) > \epsilon] \leq \mathbb{P}_{S\sim D^m}\left[b_0 < \theta_{1_0}\right] + \mathbb{P}_{S\sim D^m}\left[b_1 > \theta_{1_1}\right].$$

The event $b_0 < \theta_{1_0}$ happens if and only if all examples in *S* are not in the interval $(\theta_{1_0}, \theta_1{}^{\star})$ whose probability mass is defined to be ϵ, namely,

$$\mathbb{P}_{S\sim D^m}\left[b_0 < \theta_{1_0}\right] = \mathbb{P}_{S\sim D^m}\left[\forall(x,y) \in S, x \notin \left(\theta_{1_0}, \theta_1{}^{\star}\right)\right] = (1-\epsilon)^m \leq e^{-\epsilon m}$$

Since we assume $m > \log(2/\delta)/\epsilon$, it follows that the equation is at most $\delta/2$. In the same way it is easy to see that $\mathbb{P}_{S\sim D^m}\left[b_1 > \theta_{1_1}\right] \leq \delta/2$. In combining with the previous equation, we conclude that $\mathcal{H}$ *is PAC learnable, using the ERM algorithm* with sample complexity of $m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \log(2/\delta) / \epsilon \rceil$.

c) Give an algorithm A and determine an upper bound on the sample complexity $m_{\mathcal{H}}(\epsilon, \delta)$ such that the definition of PAC-learnability is satisfied.

Considering the realizability assumption, A hypothesis class $\mathcal{H}$ *is* PAC learnable if there exist a function $m_{\mathcal{H}}: (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: For every $\epsilon, \delta \in (0,1)$, for every distribution $\mathcal{D}$ *over* $\mathcal{X}$,and for every labeling Function $f: \mathcal{X} \to \{0,1\}$, if the realizable assumption holds with respect to $\mathcal{H}, \mathcal{D}, f$,then when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$, i.i.d. examples generated by $\mathcal{D}$ and labeled by f, the algorithm returns a hypothesis h such that, with probability of at least $1-\delta, L_{(\mathcal{D},f)}(h) \leq \epsilon$.

$m_{\mathcal{H}}: (0,1)^2 \to \mathbb{N}$ determines the *sample complexity* of learning of the hypothesis.

We know that $m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\log(|\mathcal{H}|/\delta)}{\epsilon} \right\rceil$. Knowing that PAC learnability of a class is not its finiteness but rather a combinatorial measure called the *VC dimension*.

Question – 6

A decision list may be thought of as an ordered sequence of if-then-else statements. The sequence of conditions in the decision list is tested in order, and the answer associated with the first satisfied condition is output.
More formally, a k-decision list over the boolean variables $x_1, x_{2,\dots,} x_n$ is an ordered sequence $L = \{(c_1, b_1), (c_2, b_2), \dots, (c_l, b_l)\}$and a bit b, in which each c_i is a conjunction of at most k literals over $x_1, x_2, \dots, x_n$ and each $b_i \in \{0,1\}$. For any input $a \in \{0,1\}^n$, the value $L(a)$ is defined to be b_j where j is the smallest index satisfying $c_j(a) = 1$; if no such index exists, then $L(a) = b$. Thus, b is the "default" value in case a falls off the end of the list. We call b_i the bit associated with the condition c_i.
The next figure shows an example of a 2-decision list along with its evaluation on a particular input.

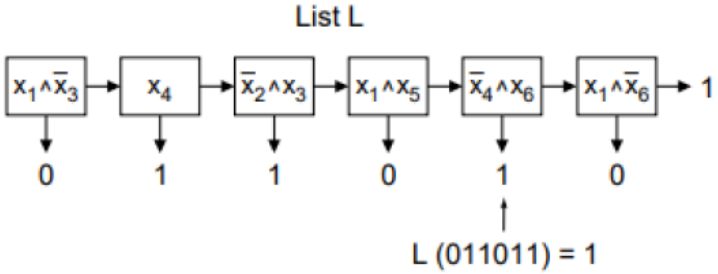


Figure 1: *A 2-decision list and the path followed by an input. Evaluation starts at the leftmost item and continues to the right until the first condition is satisfied, at which point the binary value below becomes the final result of the evaluation.*

Show that the VC dimension of 1-decision lists over $\{0,1\}^n$ is lower and upper bounded by linear functions, by showing that there exists $\alpha, \beta, \gamma, \delta \in \mathbb{R}$ such that:

$$\alpha.n + \beta \leq VC\dim(\mathcal{H}_{1-decision\ list}) \leq \gamma.n + \delta$$

Hint: Show that 1-decision lists over $\{0,1\}^n$ compute linearly separable functions (half spaces).

We will prove that the VC dimension of the class H_n of halfspaces in n dimensions is $n + 1$. While H_n is the set of functions $w_1x_1 + \cdots + w_nx_n \leq w_0$,where $w_0, \dots, w_n$ are real valued. We will use the following definition: the convex hull of a set of points S is the set of all convex combinations of points S; this is the set of all points that can be written as

$\sum_{xi \in S} \lambda_i x_i$, where $\lambda_i \leq 0$, and $\sum_i \lambda_i = 1$. It is not hard to see that if a halfspace has all points from a set S on one side, then the entire convex hull of S must be on that side as well.

1- **Upper Bound:** using Radon's theorem, let S be a set of $n + 2$ points in n dimensions. Then S can be partitioned into two (disjoint) subsets S_1 and S_2 whose convex hulls intersect. Show that Radon's Theorem implies that the VC-dimension of halfspaces is at most $n + 1$. Conclude that VC-dim (H_n) = $n + 1$.

If S is a set of $n+2$ points, then by Radon's theorem we may partition S into sets S_1 and S_2 whose convex hulls intersect. Let $p \in S_1$ be a point in that intersection. Assume there exist a hyperplane

$w.x_i \leq w_0\,, \qquad \forall x_i \in S_1$

$w.x_i > w_0\,, \qquad \forall x_i \in S_2$

So that $w.p \leq w_0$ which is contradicted by

$$w.p = \sum_{i:x_i \in S_2} \lambda_i w\,.x_i \; > \; (\sum_{i:x_i \in S_2} \lambda_i) \min_{i:x_i \in S_2} (w\,.x_i) = \min_{i:x_i \in S_2} (w\,.x_i) > w_0$$

Therefore, no set of $n+2$ points can be shattered and VC-dim (H_n) = $n+1$.

2- **Upper Bound part 2:** Prove Radon's Theorem. As a first step show the following. For a set of $n+2$ points $x_1, \dots, x_{n+2}$ in n -dimensional space, there exist $\lambda_1, \dots, \lambda_{n+2}$ not all zero such that $\sum_i \lambda_i x_i = 0$ and $\sum_i \lambda_i = 0$. (This is called affine dependence.) We first prove the affine dependence. Let $v_1, \dots, v_{n+2}$ be defined as $v_i = \langle x_i, 1\rangle \in \mathbb{R}^{n+1}$. So $\{v_i\}$ is linearly dependent. And so there exist a set of scalares $\lambda_1, \dots, \lambda_{n+2}$, not all zero, so that $\sum_i \lambda_i x_i = 0$. These $\{\lambda_i\}$ satisfy the required conditions. Further $S_1 = \{x_i | \lambda_i \geq 0\}$ and $S_2 = \{x_i | \lambda_i < 0\}$. So

$$\lambda^* = \sum_{i:x_i \in S_1} \lambda_i = \; - \sum_{j:x_j \in S_2} |\lambda_j|$$

Define

$$x^* = \sum_{i:x_i \in S_1} \lambda_i x_i = \; - \sum_{j:x_j \in S_2} |\lambda_j| x_j$$

Consider the point

$$\frac{x^*}{\lambda^*} = \sum_{i:x_i \in S_1} \frac{\lambda_i}{\lambda^*} x_i = \; - \sum_{j:x_j \in S_2} \frac{|\lambda_j|}{\lambda^*} x_j$$

This point lies in the convex hull of both S_1 and S_2.

3- **Lower Bound:** we will prove that VCdim (H_n) $\geq n+1$ by presenting a set of $n+1$ points in n-dimensional space such that one can partition that set with halfspaces in all possible ways. One good set of $n+1$ points is: The origin and all points with a 1 in one coordinate and zeros in the rest (i.e., all neighbors of the origin on the Boolean cube). Let p_i be the point with a 1 in the ith coordinate. Suppose we wish to partition this set into two pieces S_1 and S_2 with a hyperplane (and, say the origin is in S_1). Then just choose the hyperplane:

$$\sum_{\{i:p_i \in S_2\}} x_i = 1/2.$$

In Addition, the length of a decision list is bounded by the length of its input; while it may query the bits in any order, it cannot query any bit more than once. Upon querying any bit at a non-terminal step, the list may end the computation on either a 0 or a 1 and output a result of either 0

or 1. At the final step, the list may either output the value of the final bit queried or output its complement. Therefore, we find that $|C_n| \leq n!.k^n$. We also have a simple proper learning algorithm for this concept class: having seen m samples from the distribution, output a decision list that is consistent with them (if such a consistent list exists).

Furthermore, For the linear halfspaces in the plane class concept, any three points that are not collinear can be shattered. The figure (2(a)) represents how one dichotomy out of the possible 8 dichotomies can be realized by a halfspaces; the reader can easily verify that the remaining 7 dichotomies can be realized by halfspaces. To see that no set of four points can be shattered, we consider two cases. In the first case (shown in figure 2(b)), all four points lie on the convex hull defined by the four points. In this case, if we label one "diagonal" pair positive and the other "diagonal" pair negative as shown in figure (2(b)), no halfspace can induce this labeling. In the second case (shown in figure (2(c))), three of the four points define the convex hull of the four points, and if we label the interior point negative and the hull points positive, again no halfspace can induce the dichotomy. Thus, the VC dimension here is three. In general, for halfspaces in $\mathbb{R}^d$, the VC dimension is $d + 1$.

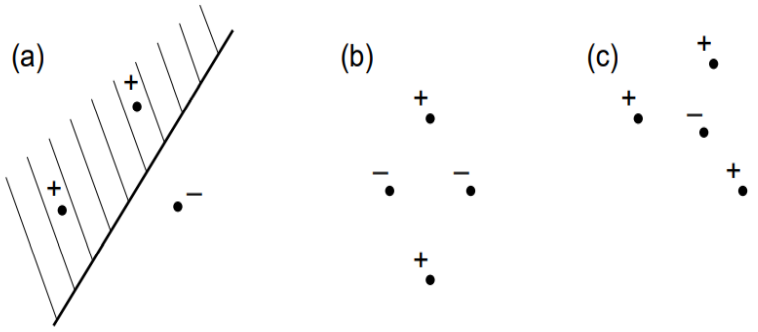


Figure 2