Assignment 1 - AML

Sîrbu Oana-Adriana, 407 AI

April 2024

1. **(0.5 points)** Give examples for:

a) a finite hypothesis class $\mathcal{H}$ with VCdim($\mathcal{H}$) = 2024. Justify your choice. **(0.25 points)**

b) an infinite hypothesis class $\mathcal{H}$ with VCdim($\mathcal{H}$) = 2024. Justify your choice. **(0.25 points)**

Solution:

a) To solve this problem I will use some information from the third Seminar.
Let $\mathcal{H}^d_{con}$ be the class of Boolean conjunctions over the variables $x_1, x_2, \dots, x_d, d \geq 2$

$$\mathcal{H}^d_{con} = \left\{ h\colon \{0,1\}^d \to \{0,1\},\ h(x_1, x_2, \dots, x_d) = \bigwedge_{i=\overline{1,d}} l(x_i) \right\}$$

$$l(x_i) = \text{literal of variable } x_i$$

$$l(x_i) \in \{x_i, \overline{x_i}, \underset{missing}{1}\}$$

We also consider that $h^- \in \mathcal{H}^d_{con}, \quad h^-(x_1, x_2, \dots, x_d) = 0$ always.
Therefore, by taking into account this definition, we can conclude that $\mathcal{H}^d_{con}$ is finite, having

$$|\mathcal{H}^d_{con}| = 3^d + 1 \tag{1}$$

Moreover, in the same seminar we proved that $\mathcal{H}^d_{con}$ shatters the set of unit vectors C $= \{e_i, i \leq d\}$ where $e_i = (0, 0, \dots, 0, \underset{i}{1}, 0, \dots, 0)$, therefore $VC\dim(\mathcal{H}^d_{con}) \geq d$ (2).

We also showed that $VC\dim(\mathcal{H}^d_{con}) < d + 1$ (3).
From (2), (3) => $VC\dim(\mathcal{H}^d_{con}) = d$, where $\mathcal{H}^d_{con}$ is finite (1).
Therefore if we take $d = 2024$, we obtain $VC\dim(\mathcal{H}^{2024}_{con}) = 2024$.

b) For this part, I have also used Seminar 3, where we defined the class of axis-aligned rectangles in $\mathbb{R}^d$ as:

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

$$\underset{\underline{x}=(x^1,x^2,\dots,x^d)}{h_{(a_1,b_1,a_2,b_2,\dots,a_d,b_d)}(\underline{x})} = \begin{cases} 1, & a_i \leq x^i \leq b_i \quad \forall i = \overline{1,d} \\ 0, & \text{otherwise} \end{cases}$$

From the definition we can see that $|\mathcal{H}^d_{rec}| = \infty$ (1).
In the seminar we proved that there exists a set C of $2d$ points that is shattered by $\mathcal{H}^d_{rec}$, which means that $VC\dim(\mathcal{H}^d_{rec}) \geq 2d$ (2).

We also showed that every set C of $2d + 1$ points is not shattered by $\mathcal{H}^d_{rec}$, which means that $VC\dim(\mathcal{H}^d_{rec}) < 2d + 1$ (3).

Therefore we have that $VC\dim(\mathcal{H}^d_{rec}) = 2d$ from (2) and (3), while we also know that the chosen hypothesis class is infinite (1).

Then, if we take $d = 1012$, we have $VC\dim(\mathcal{H}^{1012}_{rec}) = 2024$.

2. **(1 point)** 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 $= \alpha$ (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.

Solution:

I will state the definition from the lectures as a starting point for my solution.

From the definition of PAC-learnability, we know that $\mathcal{H} = \mathcal{H}_\alpha$ is PAC-learnable if there exists a function $m_\mathcal{H} \colon (0,1)^2 \to \mathbb{N}$ and there exists a learning algorithm A with the following property: for every $\epsilon, \delta > 0$, for every labeling function $f \in \mathcal{H}_\alpha$ (realizability case), for every distribution $\mathcal{D}$ on $\mathbb{R}^2$ when we run the learning algorithm A on a training set S consisting of $m \geq m_\mathcal{H}(\epsilon, \delta)$ examples sampled i.i.d. from $\mathcal{D}$ and labeled by f, the algorithm A returns a hypothesis $h_S \in \mathcal{H}$ such that, with probability at least $1 - \delta$ (over the choice of examples), the real risk of h_S is smaller than ϵ:

$$\underset{S \sim \mathcal{D}^m}{P} (L_{f,\mathcal{D}}(h_S) \leq \epsilon) \geq 1 - \delta \quad \text{which is equivalent to}$$
$$\underset{S \sim \mathcal{D}^m}{P} (L_{f,\mathcal{D}}(h_S) > \epsilon) < \delta$$

The first step is to find the algorithm A such that the class $\mathcal{H}_\alpha$ is PAC-learnable according to the theoretical definition. Because we were told to consider the realizability assumption, we know there must exist a labeling function $f \in \mathcal{H}, f = h^*_{(a_1^*, b_1^*, a_2^*, b_2^*)}$ that labels the training data. The terms $a_1^*, b_1^*, a_2^*, b_2^*$ are not randomly determined, as they form a right triangle ΔCAB which has the following property: the two catheti AB and AC parallel to the axes (Ox and Oy), and with the ratio AB/AC $= \alpha$ (fixed constant > 0). In figure 1 we can see such a representation (and with this case we will work the entire exercise), which at the same time underlies the fact that all the points inside this triangle, being labeled by h^*, will be a '+' (meaning label 1), while all the points outside this triangle will be labeled '-' (label 0).

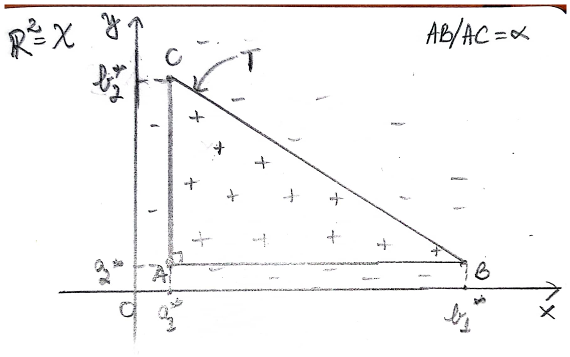

Figure 1: All the points that fall in triangle T will be labeled by h^* with label 1 (+), the other points will be labeled with label 0 (-).

Next, we will consider the training set $S = \left\{ (x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m) \;\middle|\; \begin{array}{c} y_i = h^*_{(a_1^*, b_1^*, a_2^*, b_2^*)}(x_i), \\ x_i \in \mathbb{R}^2, x_i = (x_{i1}, x_{i2}) \end{array} \right\}$

Consider the following learning algorithm A, that takes as input the training set S and outputs h_S. $h_S = h_{(a_{1s}, b_{1s}, a_{2s}, b_{2s})}$, where

$$a_{1s} = \min_{\substack{i=\overline{1,m} \\ y_i = 1}} x_{i1} \qquad\qquad a_{2s} = \min_{\substack{i=\overline{1,m} \\ y_i = 1}} x_{i2}$$

these coordinates being specifically chosen such that they represent the point in the training sample S which has mininum coordinates.

The most distant point (from the previous pair) that would define the region with the smallest area (in which we need to find all positive labels) is also important.

For $\alpha > 1$, we take for the second area-defining point the pair of coordinates (b_{1S}, a_{2S}) correspoding to the B' (starting from A' towards the hypothenuse, this is the greatest distance achievable).

$$b_{1s} = \max_{\substack{i=\overline{1,m} \\ y_i=1}} x_{i1}$$

For $\alpha < 1$, we take for the second area-defining point the pair of coordinates (a_{1S}, b_{2S}) correspoding to the C' (starting from A' towards the hypothenuse, this is the greatest distance achievable).

$$b_{2s} = \max_{\substack{i=\overline{1,m} \\ y_i=1}} x_{i2}$$

The computation for the 4th coordinate, b_{2s} (for $\alpha > 1$) or b_{1s} (for $\alpha < 1$) that would completely define the right triangle is really easy to determine using geometry. For both cases the approach is the same.

$$\frac{A'B'}{A'C'} = \alpha <=>$$

$$\frac{\sqrt{(b_{1s} - a_{1s})^2 + (a_{2s} - a_{2s})^2}}{\sqrt{(a_{1s} - a_{1s})^2 + (b_{2s} - a_{2s})^2}} = \alpha <=>$$

$$\frac{\sqrt{(b_{1s} - a_{1s})^2}}{\sqrt{(b_{2s} - a_{2s})^2}} = \alpha <=> \frac{|b_{1s} - a_{1s}|}{|b_{2s} - a_{2s}|} = \alpha$$

Because from the definition of the triangle $b_{1s} > a_{1s}$ and $b_{2s} > a_{2s}$, then the previous equation will be equivalent to:

$$\frac{b_{1s} - a_{1s}}{b_{2s} - a_{2s}} = \alpha$$

From which we obtain:

$$b_{2s} = a_{2s} + \frac{1}{\alpha}(b_{1s} - a_{1s})$$

or, if the other case is needed:

$$b_{1s} = a_{1s} + \alpha(b_{2s} - a_{2s})$$

Therefore the new hyphothesis is described by the $\Delta C'A'B'$ triangle, as presented in Figure 2.

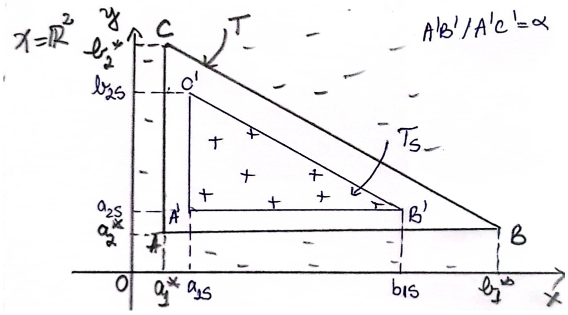

Figure 2: Triangle T_S is the tightest triangle enclosing all positive examples.

If all $y_i = 0$, then all points x_i have label 0, so there is no positive example. In this case, we choose $z = (z_1, z_2)$ a point that is not in the training set S and take $a_{1S} = b_{1S} = z_1$, $a_{2S} = b_{2S} = z_2$.

By construction, A is an ERM, meaning that $L_{h^*,\mathcal{D}}(h_S) = 0$, h_S doesn't make any errors on the training set S.

Now we want to find the sample complexity $m_{\mathcal{H}}(\epsilon, \delta)$ such that

$$\underset{S \sim \mathcal{D}^m}{P}(L_{h^*,\mathcal{D}}(h_S) \leq \epsilon) \geq 1 - \delta \text{ where } S \text{ contains } m \geq m_{\mathcal{H}}(\epsilon, \delta) \text{ examples.}$$

We make the observation that h_S makes errors in region $T \setminus T_S$, assigning the label 0 to points that should get label 1. All points $\in T_S$ will be labeled correctly (label 1), all points outside T will be labeled correctly (label 0), where T and T_S are the triangles ΔBCA and $\Delta B'C'A'$, as illustrated in the Figure 2.

We fix $\epsilon > 0, \delta > 0$ and consider a distribution $\mathcal{D}$ over $\mathbb{R}^2$.

Case i)

$$\text{If } \mathcal{D}(T) = \underset{x\sim\mathcal{D}}{P}(x \in T) \leq \epsilon \text{ then for this specific case}$$

$$L_{h^*,\mathcal{D}}(h_S) = \underset{x\sim\mathcal{D}}{P}(h_S(x) \neq h^*(x)) = \underset{x\sim\mathcal{D}}{P}(x \in T \setminus T_S) \leq \underset{x\sim\mathcal{D}}{P}(x \in T) \leq \epsilon \text{ so we have that}$$

$$\underset{S\sim\mathcal{D}^m}{P}(L_{h^*,\mathcal{D}}(h_S) \leq \epsilon) = 1 \text{ (this happens all the time)}$$

Case ii) $\mathcal{D}(T) = \underset{x\sim\mathcal{D}}{P}(x \in T) > \epsilon$

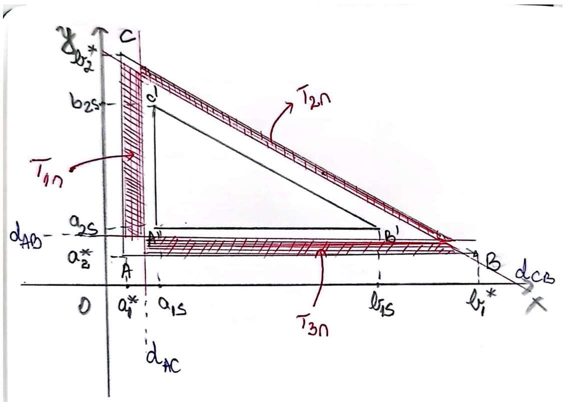

Figure 3: Constructing the triangles $T_{1\cap}$, $T_{2\cap}$ and $T_{3\cap}$.

The three geometrical figures $T_{1\cap}$, $T_{2\cap}$ and $T_{3\cap}$ were built by drawing parallel lines (d_{AB}, d_{CB} and d_{AB}) to each side of the triangle. We can see that their intersections represent some trapezoids (shaded with red). The coordinates of each trapezoid's corners can be determined using geometric formulas.

We also consider $\mathcal{D}(T_{i\cap}) = \underset{x\sim\mathcal{D}}{P}(x \in T_{i\cap}) = \frac{\epsilon}{3},\ i = \overline{1,3}$.

If T_S is the triangle returned by the learning algorithm A, implemented by h_S and it intersects each $T_{i\cap}, i = \overline{1,3}$:

$$L_{h^*,\mathcal{D}}(h_S) = \underset{x\sim\mathcal{D}}{P}(h^*(x) \neq h_S(x)) = \underset{x\sim\mathcal{D}}{P}(x \in T \setminus T_S) \leq \underset{x\sim\mathcal{D}}{P}(x \in T_{1\cap} \cup T_{2\cap} \cup T_{3\cap}) \leq$$
$$\leq \sum_{i=1}^{3} \underset{x\sim\mathcal{D}}{P}(x \in T_{i\cap}) = \sum_{i=1}^{3} \mathcal{D}(T_{i\cap}) = 3 \cdot \frac{\epsilon}{3} = \epsilon$$

So, in this case, $\underset{S\sim\mathcal{D}^m}{P}(L_{h^*,\mathcal{D}}(h_S) \leq \epsilon) = 1$ (this happens all the time).

In order to have $L_{h^*,\mathcal{D}}(h_S) > \epsilon$, we need that T_S will not intersect at least one trapezoid $T_{i\cap}$.

We denote with F_i this event, so we have $F_i = \{S \sim \mathcal{D}^m \mid T_S \cap T_{i\cap} = \emptyset\}$. This leads to the following:

$$\underset{S\sim\mathcal{D}^m}{P}(L_{h^*,\mathcal{D}}(h_S) > \epsilon) \leq \underset{S\sim\mathcal{D}^m}{P}(F1 \cup F2 \cup F3) \leq \sum_{i=1}^{3} \underset{S\sim\mathcal{D}^m}{P}(F_i)$$

where the previous inequality comes from the Union Bound theorem.

$$\begin{aligned}\underset{S\sim\mathcal{D}^m}{P}(F_i) &= \text{ what is the probability that } T_S \text{ will not intersect } T_{i\cap} \\ &= \text{ the probability that no point from } T_{i\cap} \text{ is sampled in } S \\ &= \left(1 - \frac{\epsilon}{3}\right)^m\end{aligned}$$

Therefore

$$\underset{S\sim\mathcal{D}^m}{P}(L_{h^*,\mathcal{D}}(h_S) > \epsilon) \leq \sum_{i=1}^{3} \underset{S\sim\mathcal{D}^m}{P}(F_i) = 3\cdot\left(1-\frac{\epsilon}{3}\right)^m$$

We will use the known formula:

$$1 - x \leq e^{-x}$$

to obtain

$$1 - \frac{\epsilon}{3} \leq e^{-\frac{\epsilon}{3}}$$

which helps us conclude that

$$\underset{S\sim\mathcal{D}^m}{P}(L_{h^*,\mathcal{D}}(h_S) > \epsilon) \leq 3\cdot\left(1-\frac{\epsilon}{3}\right)^m \leq 3\cdot e^{-\frac{\epsilon}{3}m}$$

Going back to the PAC learnability definition and because we already defined a $\delta > 0$, we make the previous probability very small:

$$\begin{aligned} 3\cdot e^{-\frac{\epsilon}{3}m} &< \delta \\ e^{-\frac{\epsilon}{3}m} &< \frac{\delta}{3} \quad \Big| \cdot \log_e \\ -\frac{\epsilon}{3}\cdot m &< \log\frac{\delta}{3} \quad \Big| \cdot \left(-\frac{3}{\epsilon}\right) \\ m &> -\frac{3}{\epsilon}\log\frac{\delta}{3} = \frac{3}{\epsilon}\log\frac{3}{\delta} \end{aligned}$$

Therefore, we found that $\mathcal{H}_\alpha$ is (ϵ, δ) PAC learnable by using the learning algorithm A earlier defined, while having $m \geq m_{\mathcal{H}}(\epsilon, \delta) = \frac{3}{\epsilon}\cdot\log\frac{3}{\delta}$.

3. **(1 point)** Consider $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2 \cup \mathcal{H}_3$, where:

$$\begin{aligned} \mathcal{H}_1 &= \{h_a : \mathbb{R} \to \{0,1\} \mid h_a(x) = \mathbb{1}_{[x\geq a]}(x) = \mathbb{1}_{[a,+\infty)}(x), a \in \mathbb{R}\}, \\ \mathcal{H}_2 &= \{h_b : \mathbb{R} \to \{0,1\} \mid h_b(x) = \mathbb{1}_{[x<b]}(x) = \mathbb{1}_{(-\infty,b)}(x), b \in \mathbb{R}\}, \\ \mathcal{H}_3 &= \{h_{c,d} : \mathbb{R} \to \{0,1\} \mid h_{c,d}(x) = \mathbb{1}_{[c\leq x\leq d]}(x) = \mathbb{1}_{[c,d]}(x), c, d \in \mathbb{R}\}. \end{aligned}$$

Consider the realizability assumption. Compute VCdim($\mathcal{H}$).

Solution:
I will use the information found in Lecture 6:
In order to show that the VCdim($\mathcal{H}$) is d, we need to show that:

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

I will start by exploring the behaviour of our hypothesis class $\mathcal{H}$ for a set $C_1 = \{p_1\}$ containing only one point, p_1.

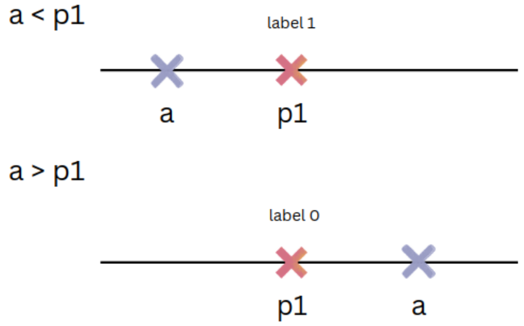

Figure 4: Representation of $\mathcal{H}_1$ shattering the set C_1

Therefore, as presented in Figure 4, if I take 2 functions ($|\mathcal{H}_{C1}| = 2^{|C|} = 2^1 = 2$) from the hypothesis class $\mathcal{H}_1$, one for which $a < p_1$ and one for which $a > p_1$, then we proved that both labelings (0 and 1) are possible (simple example: $p_1 = 1$; in the first case $a = 0$, in the second case $a = 2$). A similar behaviour is met for the classes $\mathcal{H}_2$ and $\mathcal{H}_3$, but as $\mathcal{H}$ is a reunion of all 3 classes, only proving one would suffice for the VCdim($\mathcal{H}$) ≥ 1 to be true.

Next we will consider $C_2 = \{p_1, p_2 \,|\, p_1 \leq p_2\}$. We need to have $2^2 = 4$ possible labelings: (0,0), (0,1), (1,0) and (1,1). There is no function from $\mathcal{H}_1$ that realizes the labeling (1,0) and there is no function from $\mathcal{H}_2$ that realizes the labeling (0,1), no matter how we assign the values a or b in the functions' definitions.

We will study the hypothesis class $\mathcal{H}_3$ for the same set of two points. For this part, I will also illustrate the possible cases for an easier understanding, in Figure 5.

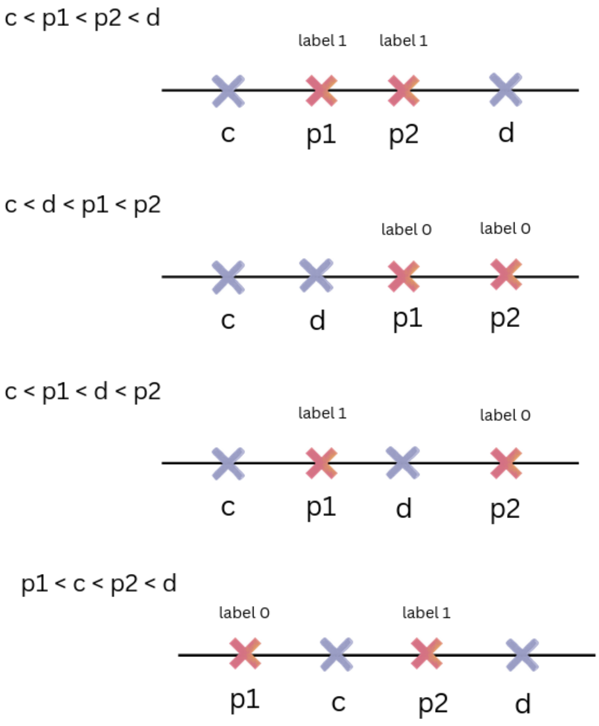

Figure 5: Representation of $\mathcal{H}_3$ shattering the set C_2

Some other configurations to obtain the same result are possible, but the important part is that we proved that $\mathcal{H}_3$ shatters the set of 2 points C_2. As $\mathcal{H}_3 \subset \mathcal{H}$ we can deduce that:

$$VCdim(\mathcal{H}) \geq 2$$

Now I will show that there is no possible way to shatter a set $C_3 = \{p_1,\, p_2,\, p_3 \;|\; p_1 < p_2 < p_3\}$ of 3 distinct points, no matter how we will use those 3 hypothesis classes that are part of $\mathcal{H}$.

We cannot find any functions h_a such that $\mathcal{H}_1$ will shatter a set of 3 distinct points. We can compute configurations to obtain:

- (1, 1, 1) - for $a < p_1 < p_2 < p_3$
- (0, 0, 0) - for $p_1 < p_2 < p_3 < a$
- (0, 1, 1) - for $p_1 < a < p_2 < p_3$
- (0, 0, 1) - for $p_1 < p_2 < a < p_3$

but we cannot compute the labelings (0, 1, 0) - because that would imply p_1 and p_3 to be smaller than a and p_2 greater than a, which is a contradiction, as we know that $p_1 < p_2 < p_3$. We can find similar contradictions for the pair of labels (1, 0, 1), (1, 1, 0), (1, 0, 0), too.

Going further, we will analyze the hypothesis class $\mathcal{H}_2$.

Using a similar approach, we prove that we can find the following labels configurations:

- (1, 1, 1) - for $p_1 < p_2 < p_3 < b$
- (0, 0, 0) - for $b < p_1 < p_2 < p_3$
- (1, 0, 0) - for $p_1 < b < p_2 < p_3$
- (1, 1, 0) - for $p_1 < p_2 < b < p_3$

The functions $h_b, \forall b \in \mathbb{R}$ don't help us compute the labelings (1, 0, 1), (0, 0, 1), (0, 1, 1) and (0, 1, 0), as we will always reach some contradictions.

From the previous 2 hypothesis classes analysed, we couldn't find a way to obtain the labels in these forms: (0, 1, 0) and (1, 0, 1).

We still have a third class to explore. For this one, using the same approach as for the rest of the problem, I will use illustrations to prove my point.

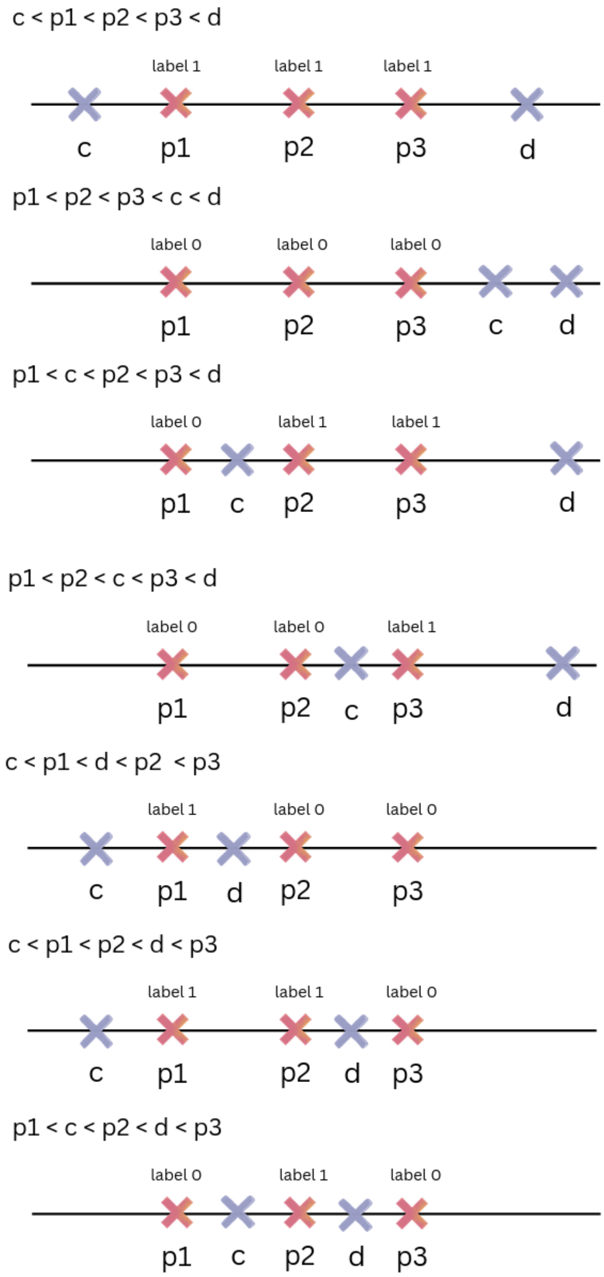

Figure 6: Representation of the 7 possible labels configurations, computed by functions $h_{c,d}$ from $\mathcal{H}_3$

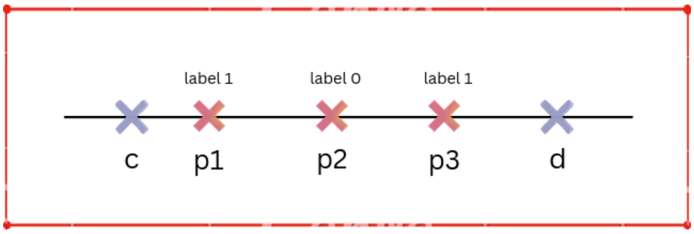

Figure 7: The contradiction that proves that $\mathcal{H}_3$ doesn't shatter the set C_3

In the previous figure is represented a contradiction. If the points p_1 and p_3 have the label 1, that means they are in the interval $[c, d]$. As $p_1 < p_2 < p_3$ that implies that p_2 is also in the interval $[c, d]$, so its label should be 1, not 0. Therefore this case is not valid.

With the help of the Figure 7 and taking into considerations all the previous statements about the hypothesis classes that are merged into $\mathcal{H}$ which can not shatter a set C_3 of 3 points as the label (1, 0, 1) is not obtained by any of the functions from those 3 classes, we proved:

- There exists a set C of size 2 that is shattered by $\mathcal{H}$. (equivalent to VCdim$(\mathcal{H}) \geq 2$).
- Every set C of size 3 is not shattered by $\mathcal{H}$. (equivalent to VCdim$(\mathcal{H}) < 2 + 1 = 3$)

In conclusion, VCdim$(\mathcal{H}) = 2$, as expected.

4. **(1 point)** Consider $\mathcal{H}$ the class of 3-piece classifiers (signed intervals):

$$\mathcal{H} = \{h_{a,b,s} : \mathbb{R} \to \{-1, 1\} \mid a \leq b, s \in \{-1, 1\}\}, \text{ where } h_{a,b,s}(x) = \begin{cases} s, & x \in [a, b] \\ -s, & x \notin [a, b] \end{cases}$$

a. Compute the shattering coefficient $\tau_H(m)$ of the growth function for $m \geq 0$ for hypothesis class $\mathcal{H}$. **(0.75 points)**

b. Compare your result with the general upper bound for the growth functions and show that $\tau_H(m)$ obtained at previous point a is not equal with the upper bound. **(0.1 points)**

c. Does there exist a hypothesis class $\mathcal{H}$ for which the shattering coefficient $\tau_H(m)$ is equal to the general upper bound (over or another domain $\mathcal{X}$)? If your answer is yes please provide an example, if your answer is no please provide a justification. **(0.15 points)**

Solution:

From the Lecture 7, we know the following definition: Let $\mathcal{H}$ be a hypothesis class. Then the growth function of $\mathcal{H}$, denoted by τ_H where $\tau_H : \mathbb{N} \to \mathbb{N}$, is defined as:

$$\tau_H(m) = max|H_C|$$

where $C \subseteq X$ and $|C| = m$. In other words, this can be stated as: $\tau_H(m)$ is the maximum number of different functions from a set C of size m to $\{0, 1\}$ that can be obtained by restricting $\mathcal{H}$ to C.

In order to compute $\tau_H(m)$, we should ask ourselves how many possible configurations for the labeling sets do we have. In the following lines, I will present my logic.

Let C be a set with m points, $C = \{c_1, c_2, \ldots, c_m\}$, with $c_1 < c_2 < \cdots < c_m$, always.

First of all we will search all possible labels for $\underline{s = 1}$.

$$\begin{aligned}
a < c_1 < c_2 < \cdots < c_m < b \qquad & \text{would generate labels } (1, 1, \ldots, 1) \\
c_1 < a < c_2 < \cdots < c_m < b \qquad & \text{would generate labels } (-1, 1, \ldots, 1) \\
\vdots \qquad & \\
c_1 < c_2 < \cdots < c_m < a < b \qquad & \text{would generate labels } (-1, -1, \ldots, -1)
\end{aligned}$$

The approach that I started with was really difficult to follow, as it didn't help me visualize all the valid cases that were left. Therefore, I tried another method, which proved to be more efficient. I categorized the labels based on n - the number of positive labels ("1").

Let us start with:

- $n = m$, then we have only 1 possible labeling:

$$(1, 1, \dots, 1)$$

- $n = m - 1$, then we can obtain:

$$\begin{gathered}(1, 1, \dots, 1, 1, -1)\\(1, 1, \dots, 1, -1, 1)\\(1, 1, \dots, -1, 1, 1)\\\vdots\\(1, -1, \dots, 1, 1, 1)\\(-1, 1, \dots, 1, 1, 1)\end{gathered}$$

  Are these all valid? The answer is no. Let's take a closer look. For any negative label inside the set, we obtain a contradiction. Let's analyze the case $(1, 1, \dots, 1, -1, 1)$ where the -1 label is (starting to count from 1) on the position $m - 1$. This would mean that the point $c_{m-1} \notin$ [a,b]. But this is contrary to the fact that the points c_{m-2} and c_m are labeled as 1, therefore being in [a, b]. And, as $c_{m-2} < c_{m-1} < c_m$ we can easily conclude that this is not a valid option. Similar observations can be made for all sets previously presented, with only 2 exceptions: $(1, 1, \dots, 1, 1, -1)$ and $(-1, 1, \dots, 1, 1, 1)$. So for the case $n = m - 1$ we have 2 possible label sets.

- $n = m - 2$ <=> only 2 points with label -1. We can theoretically obtain:

$$\begin{gathered}(1, 1, \dots, 1, -1, -1)\\(1, 1, \dots, -1, 1, -1)\\(1, 1, \dots, -1, -1, 1)\\\vdots\\(-1, -1, \dots, 1, 1, 1)\end{gathered}$$

  If at least 1 label -1 is inside the set we get similar contradictions to the ones that I previously pointed out for $n = m - 1$. Therefore, the only possible label sets for this case are: $(1, 1, \dots, 1, -1, -1), (-1, -1, \dots, 1, 1, 1), (-1, 1, \dots, 1, 1, -1)$. So for the case $n = m - 2$ we have 3 possible label sets.

- $n = m - 3$ <=> 3 points with label -1. This time we can label our points in the following way, without reaching contradictions:

  1. if the 3 negative labels are near each other:

  $$\begin{gathered}(-1, -1, -1 \dots, 1, 1, 1)\\(1, 1, 1 \dots, -1, -1, -1)\\\rightarrow \text{2 possible sets}\end{gathered}$$

  2. if the 2 of them are near each another:

  $$\begin{gathered}(-1, -1, 1 \dots, 1, -1)\\(-1, 1, 1 \dots, 1, -1, -1)\\\rightarrow \text{another 2 possible sets}\end{gathered}$$

  If at least one negative label is placed inside the set in any other configuration than those 4 previously presented we will get contradictions. So for the case $n = m - 3$ we have 4 possible label sets.

- ... the same logic is further applied ...

- n = 2 points with positive labels

$$\begin{gathered}(1,1,-1\ldots,-1,-1,-1)\\(-1,1,1,\ldots,-1,-1,-1)\\(-1,-1,1\ldots,-1,-1,-1)\\\vdots\\(-1,-1,-1\ldots,-1,1,1)\end{gathered}$$

  If we count the starting index of the positive labels pairs from the first one until the last one, we see that it ranges from 1 to $m-1$. Then we conclude that we have $m-1$ new label sets. If the labels wouldn't have adjacent positions, we will reach a contradiction. For example for the case:

$$(-1,1,-1\ldots,1,1,-1)$$

  the item on the third position is the problematic one, as its neighbours clearly are in the interval [a, b], suggesting it should also be in that interval, having label 1. The same logic applies for any other configuration similar to this one. So for the case $n=2$ we have $m-1$ possible label sets.

- $n=1$ <=> only one point labeled as 1. The possible configurations are:

$$\begin{gathered}(1,-1,-1\ldots,-1,-1,-1)\\(-1,1,-1,\ldots,-1,-1,-1)\\(-1,-1,1\ldots,-1,-1,-1)\\\vdots\\(-1,-1,-1\ldots,-1,-1,1)\end{gathered}$$

  All of them are valid cases (we can find a and b such that any point c_i will have positive label value while the other ones will be labeled by -1. So for the case $n=1$ we have m possible label sets.

- $n=0$ <=> we have all labels -1:

$$(-1,-1,-1\ldots,-1,-1,-1)$$

  So for the case $n=0$ we have 1 possible new label.

Let me make a summary up until this point:

$$\begin{gathered}n=m \implies \text{1 new label set}\\n=m-1 \implies \text{2 new label sets}\\n=m-2 \implies \text{3 new label sets}\\n=m-3 \implies \text{4 new label sets}\\\vdots\\n=2 \implies m-1 \text{ new label sets}\end{gathered}$$

Therefore, so far we have:

$$1+2+3+\cdots+(m-1)+m+1=\frac{m\cdot(m+1)}{2}+1$$

possible functions.

We are not done yet, as we still need to approach the case for which $\underline{s=-1}$.

The approach for this case will be pretty much similar to what we have seen for s = 1, as they are complementary. We need to pay attention to the signs, in order to find different label sets compared to what we have previously seen.

Now, the "positive" label will be considered -1.

The cases:

$$(-1,-1,-1\ldots,-1,-1,-1) - \text{all positives}$$

and

$$(1,1,1\ldots,1,1,1) - \text{all negatives}$$

will be neglected this time even from the start as we already have them in our valid label sets.

- $n = m-1 =>$ one label of "1". The label sets that can be computed are:

  $$(1,-1,-1\ldots,-1,-1,-1) \text{ which is valid}$$
  $$(-1,1,-1,\ldots,-1,-1,-1) \text{ which is not valid}$$
  $$(-1,-1,1\ldots,-1,-1,1) \text{ which is valid}$$

  We need to compare these to the case $s = 1$ where we had only 1 point with label 1 and see if we have any differences for the valid cases. We observe that the 2 valid label sets are found in the previous approach, too. Therefore we will pass them. For the case $n = m-1$ we have 0 new possible functions.

- $n = m-2 =>$ only two of "1" labels

  In this case we have these valid cases:

  $$(-1,-1,-1\ldots,-1,1,1)$$
  $$(1,1,-1,\ldots,-1,-1,-1)$$
  $$(1,-1,-1\ldots,-1,-1,1)$$

  from which only 1 is new, more precisely:

  $$(1,-1,-1\ldots,-1,-1,1)$$

  For the case $n = m-2$ we have 1 new possible function.

- $n = m-3 =>$ three of label "1"

  This time I will also present only the valid configurations, these being:

  $$(1,1,1,-1\ldots,-1,-1,-1)$$
  $$(-1,-1,-1,\ldots,-1,1,1,1)$$
  $$(1,1,-1\ldots,-1,-1,1)$$
  $$(1,-1,-1\ldots,-1,1,1)$$

  From these, only the last 2 are new. So, for the case $n = m-3$ we have 2 new possible functions.

- ......analysing further...

- $n = 2 =>$ only two "-1" labels in a set

  $$(-1,-1,1,\ldots,1,1,1)$$
  $$(1,-1,-1,\ldots,1,1,1)$$
  $$\vdots$$
  $$(1,1,1\ldots,-1,-1)$$

  If the positive labels don't come in pairs and one of them is inside the set we will obtain some contradictions to the functions' definitions. From all of the above, we have already taken into account:

  $$(-1,-1,1,\ldots,1,1,1)$$

  and

  $$(1,1,1,\ldots,1,-1,-1)$$

  So, for the case $n = 2$ we have $m-3$ new possible functions.

- $n = 1$ This time we have m possible functions:

$$\begin{gathered}(-1,1,1,\dots,1,1,1)\\(1,-1,1,\dots,1,1,1)\\\vdots\\(1,1,1\dots,1,-1)\end{gathered}$$

Let's not forget that we have already taken into consideration:

$$(1,1,1,\dots,1,1,-1)$$

and

$$(-1,1,1,\dots,1,1,1)$$

So, for the case $n = 1$ we have $m-2$ new possible functions.

Therefore, we now have:

$$1+2+3+\dots+(m-3)+(m-2) = \frac{(m-2)\cdot(m-1)}{2}$$

new possible functions.

The final result is:

$$\tau_m = \frac{(m-2)\cdot(m-1)}{2} + \frac{m\cdot(m+1)}{2} + 1$$

$$\tau_{\mathcal{H}}(m) = m^2 - m + 2$$

b) For this point I will use Sauer's Lemma taught in Lecture 7.

From there, we have the following statement:

Let $\mathcal{H}$ be a hypothesis class with VCdim($\mathcal{H}$) $\le d < \infty$. Then, for all m, we have that:

$$\tau_{\mathcal{H}}(m) \le \sum_{i=0}^{d} C_m^i$$

We need to compute the VCdim($\mathcal{H}$). We will use the solution from seminar 3 to prove that Vcdim($\mathcal{H}$) = 3.

Let's consider $C = \{c_1, c_2, c_3\}$ a set of 3 distinct points with $c_1 < c_2 < c_3$ (for example, take $c_1 = 0, c_2 = 1, c_3 = 2$).

To obtain labels $(-1,-1,-1)$, we can take for example $a = -1, b = -0.5$, $s = 1$

To obtain labels $(1,1,1)$, take $a = -3$, $b = 5$, $s = 1$

To obtain labels $(1,-1,-1)$, take $a = 0.5$, $b = 2.5$, $s = -1$

To obtain labels $(-1,1,1)$, take $a = 0.5$, $b = 2.5$, $s = 1$

To obtain labels $(-1,1,-1)$, take $a = 0.5$, $b = 1.5$, $s = 1$

To obtain labels $(1,-1,1)$, take $a = 0.5$, $b = 1.5$, $s = -1$

To obtain labels $(-1,-1,1)$, take $a = 1.5$, $b = 2.5$, $s = 1$

To obtain labels $(1,1,-1)$, take $a = 1.5$, $b = 2.5$, $s = -1$

So $\mathcal{H}$ shatters C, so $VC\dim(\mathcal{H}) \ge 3$.

Now, take C, a set of 4 points, $C = \{c_1, c_2, c_3, c_4\},\ c_1 \le c_2 \le c_3 \le c_4$.

Then $\mathcal{H}$ cannot realize the labels (1, -1, 1, -1).

This happens for any C. So $VC\dim(\mathcal{H}) < 4$. $\boxed{\text{So } VC\dim(\mathcal{H}) = 3}$

Therefore the general upper bound is:

$$\begin{aligned}\sum_{i=0}^{3} C_m^i &= C_m^0 + C_m^1 + C_m^2 + C_m^3 \\ &= \frac{m!}{m!\cdot 0!} + \frac{m!}{(m-1)!\cdot 1!} + \frac{m!}{(m-2)!\cdot 2!} \frac{m!}{(m-3)!\cdot 3!} \\ &= 1 + m + \frac{(m-1)m}{2} + \frac{(m-2)(m-1)m}{6} \\ &= \frac{m^3+5m+6}{6}\end{aligned}$$

Let us compare this result to the general upper bound. Coming back to the Sauer's Lemma:

$$m^2 - m + 2 \leq \frac{m^3+5m+6}{6}$$
$$0 \leq m^3 - 6m^2 + 11m + 6$$
$$0 \leq (m-3)(m-2)(m-1), m \in N$$

We can easily see that $\tau_{\mathcal{H}}(m)$ is not always equal to the general upper bound. In fact, these 2 are equal only for $m \in \{1,2,3\}$. In general, for $m > 3$, $m \in \mathbb{N}$, we have that $\tau_{\mathcal{H}}(m) < \sum_{i=0}^{3} C_m^i$. These results were obtained by using the sign chart of a function in mathematical analysis. For our case it looks like:

Table 1: Sign chart for the obtained inequality

m	0	1	2	3	...	∞
$m-1$	- - -	0	+++	+++	+++	+++
$m-2$	- - -	- - -	0	+++	+++	+++
$m-3$	- - -	- - -	- - -	0	++++	+++
$(m-3)(m-2)(m-1)$	- - -	- - 0++	+ 0- -	- 0 ++	+++	+++

c) Yes, there exists a hypothesis class $\mathcal{H}$ for which the shattering coefficient $\tau_H(m)$ is equal to the general upper bound (over or another domain $\mathcal{X}$).

Let the domain be $\mathcal{X} = [0,1]$. Take $\mathcal{H}_{threshold}$ restricted to $\mathcal{X}$

$$\mathcal{H}_{threshold,[0,1]} = \left\{ h_a \colon [0,1] \to \{0,1\},\ h_a(x) = \mathbb{1}_{[x<a]},\ a \in [0,1] \right\}$$

$$h_a(x) = \begin{cases} 1, & 0 \leq x < a \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

$VC\dim\left(\mathcal{H}_{threshold,[0,1]}\right) = 1$ (very similar proof with the one provided in lecture 6) → this result was used in the 1st exercise of seminar 3. I will also write the proof:

$\mathcal{H}$ can shatter any set of 1 point in [0,1], $C = \{c_1\}$, $0 < c_1 < 1$.

If I choose $0 \leq a < c_1 \leq 1 \implies h_a(c_1) = 0$

If I choose $0 \leq c_1 < a \leq 1 \implies h_a(c_1) = 1$

We should also prove that any C of 2 points, $C = \{c_1, c_2\}$ is not shattered by my $\mathcal{H}$. We cannot find a parameter a such that $h_a(c_1) = 0, h_a(c_2) = 1$.

If $a \in [0,1]$ such that $h_a(c_2) = 1 \implies c_2 < a$. But we know that $c_1 < c_2$, therefore $c_1 < a$, so $h_a(c_1)$ should be 1, not 0. Therefore we reached a contradiction.

$\implies$ VCdim($\mathcal{H} = 1$)

Knowing d = 1, we compute the general upper bound for this case.

$$\begin{aligned}\sum_{i=0}^{1} C_m^i &= C_m^0 + C_m^1 = \\ &= \frac{m!}{m!\cdot 0!} + \frac{m!}{(m-1)!\cdot 1!} = \\ &= 1 + m\end{aligned}$$

We also need to compute $\tau_{\mathcal{H}}(m)$ of the growth function for this hypothesis class.

We try to find all possible labels that the functions from $\mathcal{H}$ can generate over a set $C = \{c_1, c_2, \ldots, c_m\}$ with $0 \leq c_1 < c_2 < \cdots < c_m \leq 1$.

We can have:

$$\begin{aligned}
0 \leq c_1 < c_2 < \cdots < c_m < a \leq 1 &\text{ for which the labels are } (1,1,\ldots,1,1)\\
0 \leq c_1 < c_2 < \cdots < a < c_m \leq 1 &\text{ for which the labels are } (1,1,\ldots,1,0)\\
&\vdots\\
0 \leq c_1 < c_2 < a < \cdots < c_m \leq 1 &\text{ for which the labels are } (1,1,0,\ldots,0)\\
0 \leq c_1 < a < c_2 < \cdots < c_m \leq 1 &\text{ for which the labels are } (1,0,0,\ldots,0)\\
0 \leq a < c_1 < c_2 < \cdots < c_m \leq 1 &\text{ for which the labels are } (0,0,0,\ldots,0)
\end{aligned}$$

All of the above are valid. Paying attention to the definition of the functions from the class $\mathcal{H}$, we can't have points with label '0' before the ones with label '1', because that would be a contradictory situation, so there are not any other possible functions.

$\implies \tau_{\mathcal{H}}(m) = m + 1$

We obtained that $\tau_{\mathcal{H}}(m) = \sum_{i=0}^{1} C_m^i = m + 1$, so we proved that there exists a hypothesis class $\mathcal{H}$ for which the shattering coefficient is equal to the general upper bound.

5. **(1 point)** Let $\mathcal{H} = \{h : \mathbb{R} \to \{0,1\} \mid h_\theta(x) = \mathbb{1}_{[\theta,\theta+1]\cup[\theta+2,\infty)}(x), \theta \in \mathbb{R}\}$.
Compute VCdim($\mathcal{H}$).

Solution:

I will use once again the information found in Lecture 6:

In order to show that the VCdim($\mathcal{H}$) is d, we need to show that:

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

I will start by exploring the behaviour of our hypothesis class $\mathcal{H}$ for a set $C = \{p_1, p_2\}$ containing only two points, p_1 and p_2, with $p_1 < p_2$. For an easier understanding I will also provide some examples containing numerical values to have a full proof that such labels exist.

If C is a set of 2 points, then $|C| = 2$ and we need to have $2^{|C|} = 2^2 = 4$ functions from $\mathcal{H}$ that would make these 4 label sets possible: (0,0), (1,1), (1,0) and (0,1).

Let $p_1 = 1$ and $p_2 = 1.5$:

- If we have $\theta \leq p_1 < p_2 \leq \theta + 1 < \theta + 2 \implies$ we can reach the pair of labels (1,1). Example: $\theta = 0.9 \implies 0.9 \leq p_1 = 1 < p_2 = 1.5 \leq 1.9 < 2.9$, which is true.
- If we have $\theta \leq p_1 \leq \theta + 1 < p_2 < \theta + 2 \implies$ we can reach the pair of labels (1,0). Example: $\theta = 0.3 \implies 0.3 \leq= p_1 = 1 \leq 1.3 < p_2 = 1.5 < 2.3$, which is also true.
- If we have $p_1 < \theta \leq p_2 \leq \theta + 1 < \theta + 2 \implies$ we can reach the pair of labels (0,1). Example: $\theta = 1.25 \implies p_1 = 1 \leq 1.25 \leq p_2 = 1.5 < 2.25 < 3.25$, which is also true.
- If we have $p_1 < p_2 < \theta < \theta + 1 < \theta + 2 \implies$ we can reach the pair of labels (0,0). Example: $\theta = 2 \implies p_1 = 1 < p_2 = 1.5 < 2 < 3 < 4$, which is true.

From all the above we conclude that VCdim($\mathcal{H}$) ≥ 2.

We now take a set C of 3 points, $C = \{p_1, p_2, p_3\}$ with $p_1 < p_2 < p_3$. Because $|C| = 3$, now we need to find $2^3 = 8$ possible label sets provided by functions from the hypothesis class $\mathcal{H}$. For this part I will attach graphical representations for an easier understanding. Let's also take

$$p_1 = 0.7, p_2 = 1.25, p_3 = 1.8$$

in order to provide some numerical values to make the proof valid.

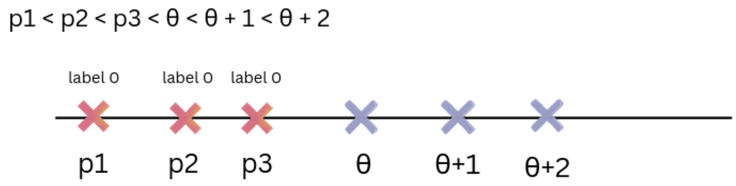

Figure 8: Having this configuration, if we take for example $\theta = 2$, we obtain labels (0,0,0)

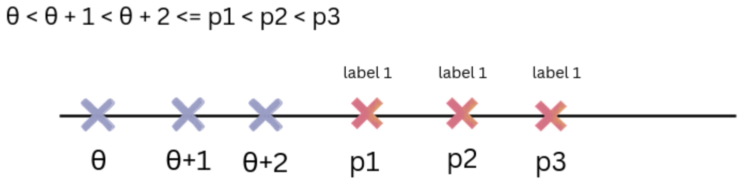

Figure 9: Having this configuration, if we take for example $\theta = -1.5$, we obtain labels (1,1,1)

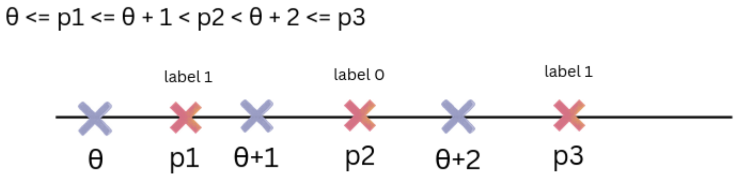

Figure 10: Having this configuration, if we take for example $\theta = -0.25$, we obtain labels (1,0,1)

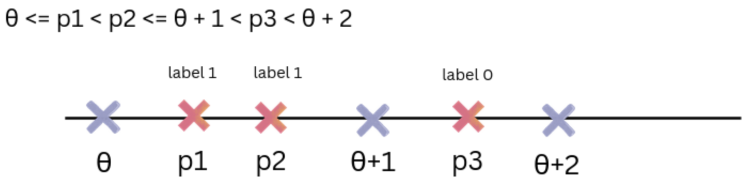

Figure 11: Having this configuration, if we take for example $\theta = 0.5$, we obtain labels (1,1,0)

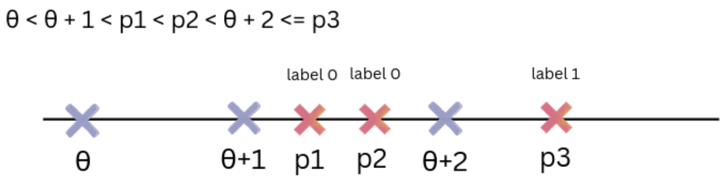

Figure 12: Having this configuration, if we take for example $\theta = -0.5$, we obtain labels (0,0,1)

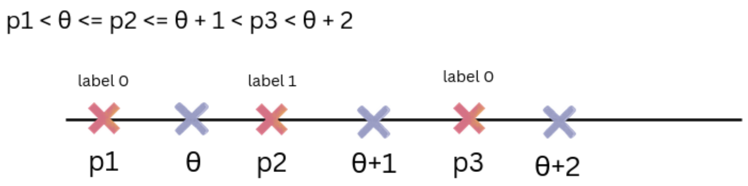

Figure 13: Having this configuration, if we take for example $\theta = 0.75$, we obtain labels (0,1,0)

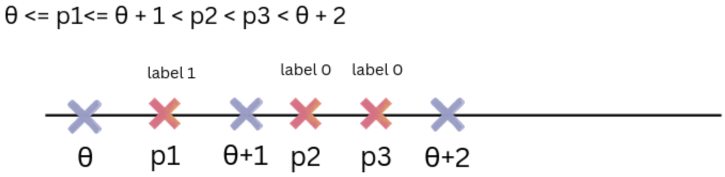

Figure 14: Having this configuration, if we take for example $\theta = -0.1$, we obtain labels (1,0,0)

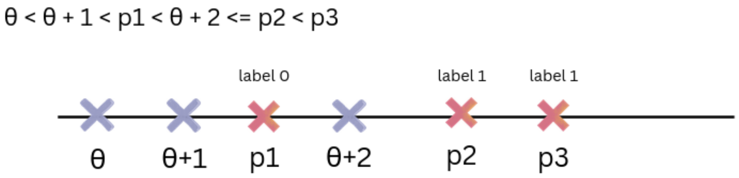

Figure 15: Having this configuration, if we take for example $\theta = -1$, we obtain labels (0,1,1)

All $2^3 = 8$ possible label sets can be achieved by using functions from $\mathcal{H}$.

$$\implies VCdim(\mathcal{H}) \geq 3$$

Now let's take $C = \{p_1, p_2, p_3, p_4\}$, with $p_1 < p_2 < p_3 < p_4$. If VCdim($\mathcal{H} = 4$) it would imply that we can find $2^{|C|} = 2^4 = 16$ label sets.

For the labels (0,0,0) and (1,1,1) it is really straightforward to find a solution as we can simply take any θ such that $p_1 < p_2 < p_3 < p_4 < \theta$ to get (0,0,0) and any θ such that $\theta + 2 < c_1 < c_2 < c_3 < c_4$ to get (1,1,1).

As we have seen in the seminar, usually the biggest problems appear for sets of alternating labels, such as (1,0,1,0). Let's try to find a function that would label our set of 4 points this way.

If p_1 has label 1 $\implies p_1 \in [\theta, \theta + 1] \cup [\theta + 2, +\infty)$. But p_1 cannot be in $[\theta + 2, +\infty)$, because there wouldn't exist a point with a value bigger than p_1 that would be labeled as 0. $\implies p_1 \in [\theta, \theta + 1]$.

If p_2 has label 0 and, as $p_2 > p_1$, p_2 must be in $(\theta + 1, \theta + 2)$.

We have $p_3 > p_2 > p_1$, with the label of p_3 being 1. This implies that $p_3 \in [\theta + 2, +\infty)$.

But we have that $p_4 > p_3 > p_2 > p_1$, the label of p_4 being 0. There is not any possible configuration to produce such a label following the logical approach that we have started with.

Therefore,

$$VCdim(\mathcal{H}) < 4$$

And, as we proved that

$$VCdim(\mathcal{H}) \geq 3$$

We conclude that

$$VCdim(\mathcal{H}) = 3$$

6. **(1 point)** 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 $x_1, x_2, \ldots, x_n$ is an ordered sequence $L = \{(c_1, b_1), (c_2, b_2), \ldots, (c_l, b_l)\}$ and a bit b, in which each c_i is a conjunction of at most k literals over $x_1, x_2, \ldots, x_n$ 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.

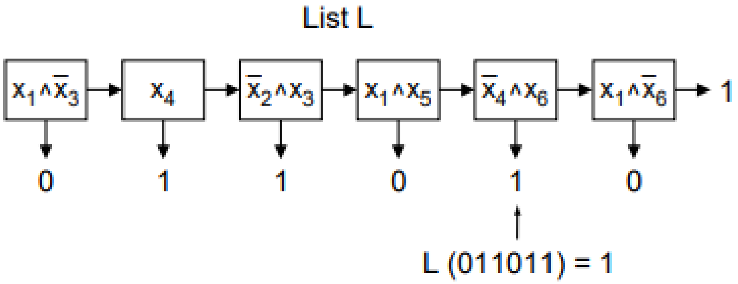

Figure 16: *A 2-decision list and the path followed by an input. Evaluation starts at the leftmost item and continues to the right until the first condition is satisfied, at which point the binary value below becomes the final result of the evaluation.*

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:

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

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

Solution:

From Lecture 8, we know that the VC dimension determines (along with ϵ, δ) the samples complexities of a learning class. It gives us a lower and upper bound. To better understand the problem, I have read a few scientific papers covering the topic, from which I will inspire when writing my solution. They will be referenced.

Let decision list = DL.

Let $L_n = \{x_1, \bar{x_1}, \ldots, x_n, \bar{x_n}\}$ denote the set of $2n$ literals associated with variables in $V_n, V_n = \{x_1, x_2, \ldots, x_n\}$ [2].

From [2] we know that, if D_k^n denotes the set of all clauses (disjunctions) of size k with literals drawn from L_n and C_k^n denote the set of all conjunctions of size at most k with literals drawn from the same set, then we have:

$$|C_k^n| = |D_k^n| = \sum_{i=0}^{k} \binom{2n}{i} = O(n^k)$$

so for $k = 1$ they have dimensions that are linear functions of n. Why is this aspect important? Because, as stated in the same paper [2] in Theorem 1 from section 3.1, for $0 < k < n$, k-CNF(n) and k-DNF(n) are proper subsets of k-DL(n). From this we deduce that:

$$VCdim(kDL) \geq VCdim(kDNF) = O(n^k)$$

and for $k = 1$:

$$VCdim(DL) \geq VCdim(DNF) = O(n)$$

therefore the lower bound is a linear function $\implies \exists\ \alpha, \beta$ such that:

$$\alpha \cdot n + \beta \leq VCdim(\mathcal{H}_{1-DL})$$

Now, to find an upper bound, we follow the provided hint and we try to prove that the class of 1DL functions is linearly separable $\Leftrightarrow$ we have a linear function:

$$P_L(x_1, \ldots, x_n) = a_1x_1 + a_2x_2 + \cdots + a_nx_n$$

having coefficients $a_i, 1 \leq i \leq n$ for each ordered sequence $L \in 1DL$. The proof is inspired by [1].

Let us have also a threshold value $\tau \in \mathbb{R}$ such that for any $x_i \in \{0, 1\}$ with $1 \leq i \leq n$:

$$P_L(x_1, \ldots, x_n) \geq \tau \Leftrightarrow L(x_1, \ldots, x_n) = 1 \tag{2}$$

I have to show that for any $L \in 1DL$, we can construct a linear function P_L that satisfies the above condition.

How do we construct P_L? For each pair (c_i, b_i) in L, we associate a linear term T_i defined by:

- $T(x_i, 1) = x_i$
- $T(x_i, 0) = 1 - x_i$
- $T(\bar{x}_i, 1) = 1 - x_i$
- $T(\bar{x}_i, 0) = x_i$

and then we construct the linear function $P_L(x_1, \ldots, x_n)$ as a linear combination of these terms. The corresponding mathematical form is:

$$P_L(x_1, \ldots, x_n) = d_1T_1 + \cdots + d_mT_m \geq \tau \tag{3}$$

with $d_i \in \mathbb{R}, 1 \leq i \leq m$ such that they satisfy equation 2. But how are the coefficients d_i and τ chosen? The answer is that they must be constructed according to some inductive rules:

1. $d_m = 1$
2. If $b_i = 0$, then $d_i = 1 + \sum_{j=i+1}^{m} d_j$
3. If $b_i = 1$ then $d_i = \sum_{j=i+1}^{k} d_j$, where k is the least l, $i + 1 \leq l \leq m$ such that $b_l = 1$ (according to the definition).
4. $\tau = \sum_{j=1}^{k} d_j$, where k is the least $l, 1 \leq l \leq m$ such that $b_l = 1$.

We assumed without loss of generality that $b_m = 1$.

If we analyze the base case for the inductive rules, more precisely if we have $m = 1$, then we have 2 possibilities:

$$L =< (x_i, 1) >, \ L =< (\bar{x}_i, 1) >$$

The linear inequalities constructed for each one of the above cases satisfy the condition 2, as they are $x_i \geq 1$ and $1 - x_i \geq 1$. Therefore, this is a valid base case.

If we have $L' =< (c_2, b_2), \ldots, (c_m, b_m) >$ (we eliminated the first item from L), by using the inductive hypothesis there must be a linear function $P_{L'}(x_1, \ldots, x_n)$ and a threshold τ' constructed by the above rules, satisfying:

$$P_{L'}(x_1, \ldots, x_n) \geq \tau' \Leftrightarrow L'(x_1, \ldots, x_n) = 1 \tag{4}$$

This result is proved in [1], therefore we conclude that 1-DL are linearly separable $\implies$

$$VCdim(1DL) = O(n)$$

so there must exist γ, δ such that:

$$VCdim(\mathcal{H}_{1-DL}) \leq \gamma \cdot n + \delta$$

Therefore, the VC dimension of 1-decision lists over $\{0, 1\}^n$ is lower and upper bounded by linear functions:

$$\alpha \cdot n + \beta \leq VCdim(\mathcal{H}_{1-DL}) \leq \gamma \cdot n + \delta$$

References

[1] Andrzej Ehrenfeucht, David Haussler, Michael Kearns, and Leslie G. Valiant. A general lower bound on the number of examples needed for learning. *Inf. Comput.*, 82:247–261, 1988.

[2] Ronald L. Rivest. Learning decision lists. *Mach. Learn.*, 2(3):229–246, nov 1987.