Assigment I

Advance Machine Learning

Dăscălescu Dana

June 12, 2022

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

Exemple 1: Let us consider $\mathcal{X}$ to be the Boolean hypercube $\{0,1\}^n$ and

$$\mathcal{H}_{n-parity} = \left\{ h_I \;\middle|\; \begin{array}{l} I \subseteq \{1,2,\ldots,n\}, x = (x_1, x_2, \ldots, x_n) \in \{0,1\}^n, \\ h_I : I \subseteq \{1,2,\ldots,n\} \to \{0,1\}, h_I(x_1, x_2, \ldots, x_n) = \left(\sum_{i \in I} x_i\right) \bmod 2 \end{array} \right\}$$

We will show that $VCdim(\mathcal{H}) = n$ and we will pick $n = 2022$ for our example.

Proof. For each subset $I \subseteq \{1,2,\ldots,n\}$, where $n = 2022$, we have a parity function h_I, so $|\mathcal{H}_{2022-parity}| = 2^{2022}$ (we also have that $\mathcal{H}_{2022-parity}$ is a finite hypotheses class).

We know that for a finite hypotheses class $\mathcal{H}$, $VCdim(\mathcal{H}) \leq \lfloor log\left(|\mathcal{H}|\right) \rfloor$. For our particular case, we have the upper bound

$$VCdim(\mathcal{H}_{2022-parity}) \leq \left\lfloor log_2\left(\left|2^{2022}\right|\right)\right\rfloor \Leftrightarrow VCdim(\mathcal{H}_{2022-parity}) \leq 2022 \tag{1}$$

Now, we will show that $VCdim(\mathcal{H}_{2022-parity}) \geq 2022$ by finding a set of 2022 entries from the Boolean hypercube $\{0,1\}^{2022}$ that is shattered by $\mathcal{H}_{2022-parity}$.

Let us consider the set of unit vectors $C = \{e_i = (0, \ldots, 0, \underset{i}{1}, 0, \ldots, 0) \mid i = \overline{1, 2022}\}$. We need to show that, for each possible labeling $(y_1, y_2, \ldots, y_{2022})$ of $(e_1, e_2, \ldots, e_{2022})$, we can find a corresponding h, $h \in \mathcal{H}_{2022-parity}$, such that $h(e_i) = y_i, \forall i = \overline{1, 2022}$.

Consider the vector of labels $(y_1, y_2, \ldots, y_{2022})$ and let $I = \{i \in [2022] : y_i = 1\}$. Then we have

$$h_I(x) = \left(\sum_{i \in I} x_i\right) \bmod 2 = \left\{ \begin{array}{l} 1, \text{ if } i \in I \\ 0, \text{ otherwise }. \end{array} \right.$$

Thus, we have $h_I(e_i) = y_i$ for every $i \in [n]$. So $VCdim(\mathcal{H}_{2022-parity}) \geq 2022$. Combining with Equation 1, we conclude our proof that $VCdim(\mathcal{H}_{2022-parity}) = 2022$.

□

Exemple 2: Consider the class $\mathcal{H}_{mcon}^{2022}$ of monotone Boolean conjunctions over $\{0,1\}^{2022}$.

$$\mathcal{H}_{mcon}^{2022} = \left\{h : \{0,1\}^{2022} \to \{0,1\}, h(x_1, x_2, \ldots, x_{2022}) = \bigwedge_{i=1}^{2022} l(x_i)\right\} \cup \{h^-\}$$

$$l(x_i) \in \{x_i, 1\}$$

In the following, we will show that $VCdim(\mathcal{H}_{mcon}^{2022}) = 2022$.

Proof. We know that $|\mathcal{H}_{mcon}^{2022}| = 2^{2022} + 1$. We can use the property of a finite hypothesis class that $VCdim(\mathcal{H}) \leq \lfloor log_2 (|\mathcal{H}|) \rfloor$. Thus, we have the upper bound $VCdim(\mathcal{H}_{mcon}^{2022}) \leq 2022$.

Subsequently, we would like to show that there exists a set $C \subseteq \{0,1\}^{2022}$ with 2022 points that is shattered by $\mathcal{H}_{mcon}^{2022}$.

We choose $C = \{(0,1,1,\ldots,1,1), \ldots, (1,\ldots,1,\underset{i}{0},1,\ldots,1), \ldots, (1,1,1,\ldots,1,0)\}$ set of vectors of the form $c_i = (1,1,\ldots,1) - e_i, i = \overline{1,2022}$, where e_i is the unit vector $(0,\ldots,0,\underset{i}{1},0,\ldots,0)$.

Consider the vector of labels $(y_1, y_2, \ldots, y_{2022})$ and let $\mathcal{I} = \{i \in [2022] : y_i = 1\}$. We want to show that there exists a function h, $h \in \mathcal{H}_{mcon}^{2022}$, such that $h(c_i) = y_i, \forall i = \overline{1,2022}$

If $\mathcal{I} = \emptyset$, then h^- realizes the labelling $\underbrace{(0,\ldots,0)}_{\text{times } 2022}$.

If $\mathcal{I} = \{1,2,\ldots,2022\}$, them h_{empty} realizes the labelling $\underbrace{(1,\ldots,1)}_{\text{times } 2022}$.

If $1 \leq |\mathcal{I}| \leq 2021$, then consider $h_{\mathcal{I}}(x_1,\ldots,x_{2022}) = \bigwedge_{i \notin \mathcal{I}} x_i$. In this case, we have that $h_{\mathcal{I}}(c_i) = 1$ if $i \in \mathcal{I}$, and $h_{\mathcal{I}}(c_i) = 0$ if $i \notin \mathcal{I}$.

For all indices $i \in \mathcal{I}$, c_i will have value 0 on the position i and 1 in rest, but variable x_i is not considered in the conjunction. So $h_{\mathcal{I}}(c_i) = 1$.

For all indices $i \notin \mathcal{I}$, c_i will have value 0 on the position i and, because the conjunction contains the literal x_i, then we have that $h_{\mathcal{I}}(c_i) = 0$.

We have that $\mathcal{H}_{mcon}^{2022}$ shatters C, so $VCdim(\mathcal{H}_{mcon}^{2022}) \geq 2022$. Thus, combining it with the upper bound, $VCdim(\mathcal{H}_{mcon}^{2022}) \leq 2022$, we concluded our proof that $VCdim(\mathcal{H}_{mcon}^{2022}) = 2022$.

$\square$

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 $m = \frac{n}{2}$? Give an example of such an $\mathcal{H}$ and C. Justify your answer.

Proof. According to the problem statement, we have $|\mathcal{H}| = n = 2m$ and $VCdim(\mathcal{H}) \geq m = \frac{n}{2}$. We also know that $VCdim(\mathcal{H}) \leq log(|\mathcal{H}|)$. Thus, we have

$$\left.\begin{array}{c} m \leq log_2(2 \cdot m) \\ m \in \mathbb{N} \end{array}\right\} \Rightarrow m \in \{1, 2\} \; .$$

Since $n = 2 \cdot m$ is the maximum value of the natural even number with the property described above, we have that $m = 2$, and hence $n = 4$.

Now, we will take the same example as in the previous exercise ($\mathcal{H}_{n-parity}$, $|\mathcal{H}_{n-parity}| = 2^n$ and $VCdim(\mathcal{H}_{n-parity}) = n$), the hypothesis class $\mathcal{H}_{2-parity}$ which has $|\mathcal{H}_{2-parity}| = 2^2 = 4$ and shatters the standard basis $\{e_j\}_{j=1}^2$.

□

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

Proof. We shall show in the following that $VCdim(\mathcal{H}) = 2$. To prove this we need to find a set of 2 points that are shattered by $\mathcal{H}$, and show that no set of 3 points can be shattered by $\mathcal{H}$.

It is easy to see that 2 points can be shattered by axis aligned rectangles with the center in the origin $O(0,0)$. Let us consider $p_1 = (x_{p_1}, y_{p_1}) \in \mathbb{R}^2$ and $p_2 = (x_{p_2}, y_{p_2}) \in \mathbb{R}^2$, such that $x_{p_1} < x_{p_2}$, $y_{p_1} > y_{p_2}$, and $C = \{p_1, p_2\}$. For this two points we can achieve all possible labelling, more precisely $(0,0), (0,1), (1,0), (1,1)$, as shown in Figure 1.

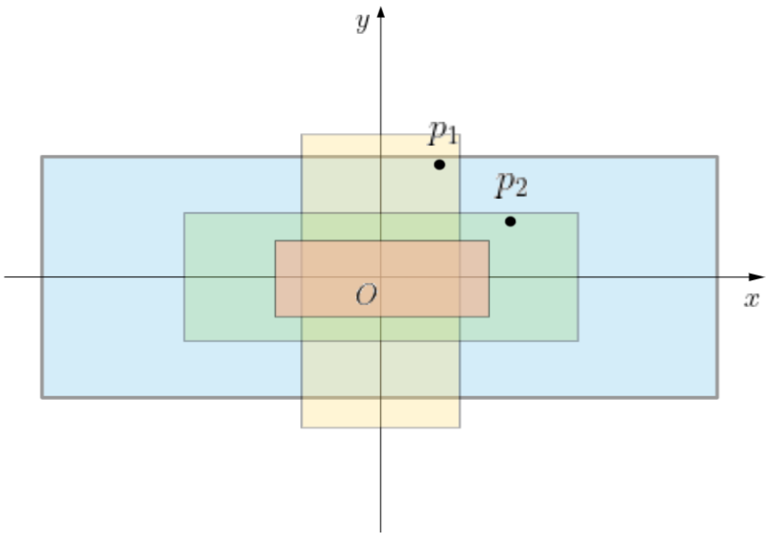


Figure 1: Two points shattered by the class of axis-aligned rectangles with the center in origin $O(0,0)$

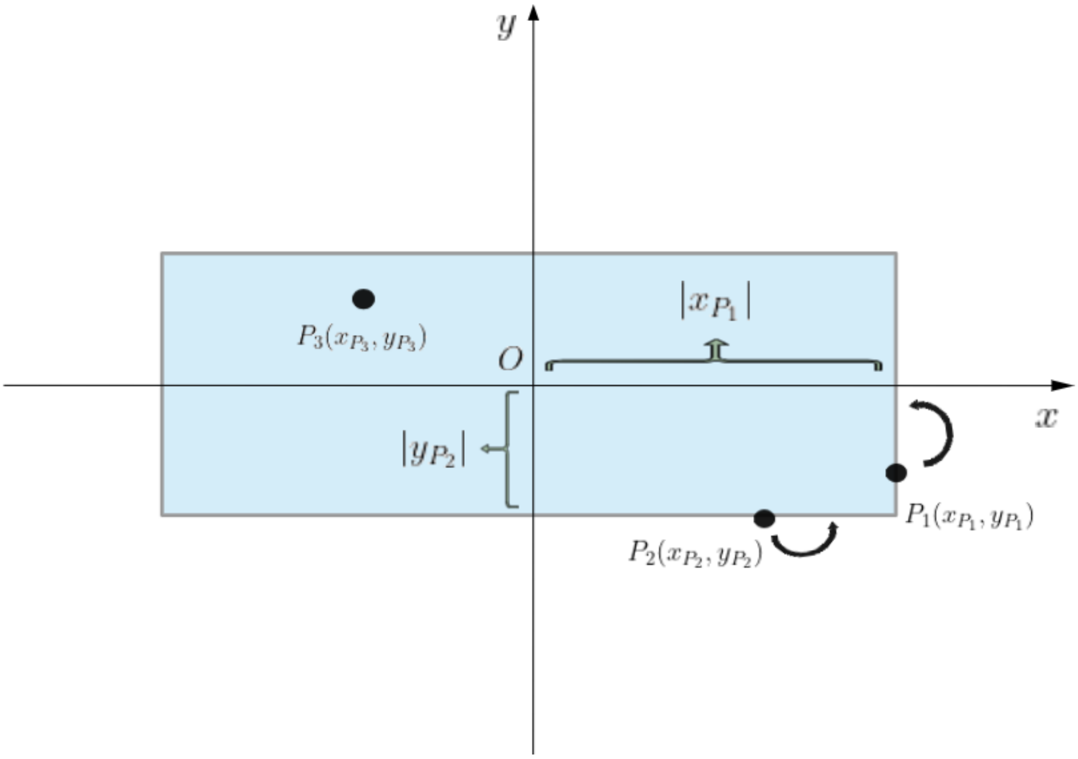


Figure 2: An impossible assigment of $+/-$ to the data, as all rectangles that contains the two points P_1 and P_2 (marked $+$) must also contain the one marked with $-$ point, P_3

$\mathcal{H}$ shatters C, so $VCdim(\mathcal{H}) \geq 2$ $(*)$.

Consider any set of three distinct points $\{p_1 = (x_{p_1}, y_{p_1}), p_2 = (x_{p_2}, y_{p_2}), p_3 = (x_{p_3}, y_{p_3})\} \subset \mathbb{R}^2$ (provided they are not collinear). These three points denote a non-degenerate triangle. Consider a rectangle centered in origin $O(0,0)$ that contains the points with maximum x-coordinate in absolute value ($\exists j_1 \in [3]$ such that $|x_{p_{j_1}}| \geq |x_{p_j}|, \forall j = \overline{1,3}$) and maximum y-coordinate in absolute value ($\exists j_2 \in [3]$ such that $|y_{p_{j_2}}| \geq |y_{p_j}|, \forall j = \overline{1,3}$). These points may not be distinct. However, there are at most two such points. Call this set of points $V \subset \{p_1, p_2, p_3\}$. There is at least one point p_i that is not in V, but still must be in the area inside of the rectangle (as shown in Figure 2). Therefore, the labelling that labels all vertices in V with $+$ and the vertice p_i with $-$ cannot be consistent with any axis aligned rectangles with center in origin $O(0,0)$. This means that there is no shattered set of size 3, since all possible labellings of a shattered set must be realized by some concept. So, $VCdim(H) < 3$ $(**)$.

$$(*)(**) \implies VCdim(H) = 2$$

$\square$

4. Axis-aligned rectangled triangles:

$$\mathcal{H}_\alpha = \left\{ \begin{array}{l} h_{\triangle ABC} \colon \mathbb{R}^2 \to \{0,1\}, \frac{AB}{AC} = \alpha, \alpha > 0, AB \parallel Ox \text{ and } AC \parallel Oy, \\ h_{\triangle ABC}(x_1, x_2) = \mathbb{1}_{\triangle ABC} \end{array} \right\}$$

Show that the class $\mathcal{H}_\alpha$ is $(\varepsilon, \delta)-$PAC learnable by giving an algorithm $\mathcal{A}$ and determing an upper bound on the sample complexity $m_H(\varepsilon, \delta)$ such that the definition of PAC-learnability is satisfied.

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

$$\underset{S \sim \mathcal{D}^m}{\mathbb{P}}[L_{D,f}(h_S) \leq \varepsilon] \geq 1 - \delta \Leftrightarrow \underset{S \sim \mathcal{D}^m}{\mathbb{P}}[L_{D,f}(h_S) > \varepsilon] < \delta$$

Firstly, we will define the algorithm $\mathcal{A}$.

We are under the realizability assumption, so there exists a function $f \in \mathcal{H}$, $f = h_{\triangle A^*B^*C^*}$ that labels the training data, where $h_{\triangle A^*B^*C^*}$ is denoted as the hypothesis represented by a orthogonal triangle $\triangle A^*B^*C^*$ with the two catheti A^*B^* and A^*C^* parallel to the axes (Ox and Oy), with the ratio $\dfrac{A^*B^*}{A^*C^*} = \alpha$, $\alpha \in \mathbb{R}^*_+$ (fixed constant), and a_1^*, b_1^*, a_2^*, b_2^* represent the coordinates of the rectangled triangle's vertices, as follows:

$$A^*(a_1^*, b_1^*), \quad B^*(a_2^*, b_1^*), \quad C^*(a_1^*, b_2^*)$$

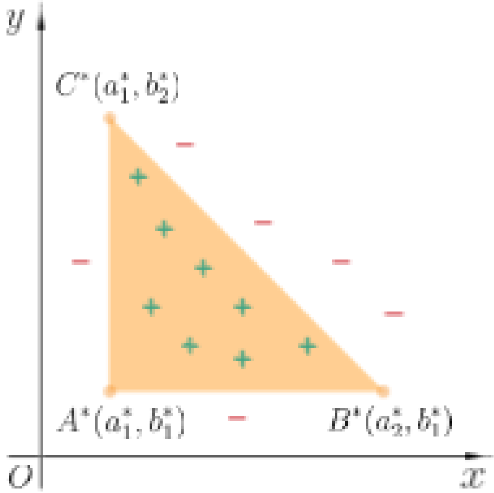


Figure 3: All the points inside the area of the right triangle $\triangle A^*B^*C^*$ will be labeled by h^* with label $1(+)$, and all the points outside will be labeled with label $0(-)$.

As in Figure 3, $h_{\triangle A^*B^*C^*}$ labels each point drawn from the rectangled triangle $\triangle A^*B^*C^*$ with label $1(+)$, all the other points with label $0(-)$.

Consider the training set $S = \left\{ (x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m) \;\middle|\; \begin{array}{l} x_i \in \mathbb{R}^2, x_i = (x_{i1}, x_{i2}), \\ y_i = h^*_{\triangle A^*B^*C^*}(x_i) \end{array} \right\}$.

Consider the following algorithm $\mathcal{A}$, that takes as input the samples from the training set and outputs a hypothesis $h_S = h_{\triangle A_S B_S C_S}$:

STEP 1: We choose a_{1S} to be the leftmost coordinate on the Ox axis of a positive point in the training set S, and b_{1S} to be the lowest coordinate on Oy axis of a positive point in the training set S:

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

We take A_S as the point with coordinates (a_{1S}, b_{1S}).

STEP 2: We need to compute B_S and C_S such that the resulting hypothesis will be a function from the hypothesis class H_α, meaning that the $\triangle A_S B_S C_S$ must have the following property: the two catheti $A_S B_S$ and $A_S C_S$ are parallel to the axes (Ox and Oy), with the ratio $\frac{A_S B_S}{A_S C_S} = \alpha$, $\alpha > 0$ (fixed constant).

Let d be an oblique line passing through the points B and C, and θ the angle formed by the line with the axis of the abscissa (see Figure 4). The real number $\tan(\pi - \theta)$ is called the slope of line d or the angular coefficient of a line (angle of inclination $\pi - \theta$) and it is indicated by m.

We consider the general equation given for a line:

$$y = mx + n$$

We know that $\tan(\theta) = \frac{AC}{AB} = \frac{1}{\alpha}$. The slope of the hypotenuse BC is given by $\tan(\pi - \theta) = \dfrac{\tan(\pi) - \tan(\theta)}{1 + \tan(\pi)\tan(\theta)} = -\tan(\theta) = -\dfrac{1}{\alpha}$. Therefore, the lines which are parallel to our hypotenuse will be given by the following equation: $-\dfrac{x}{\alpha} - y + n = 0$.

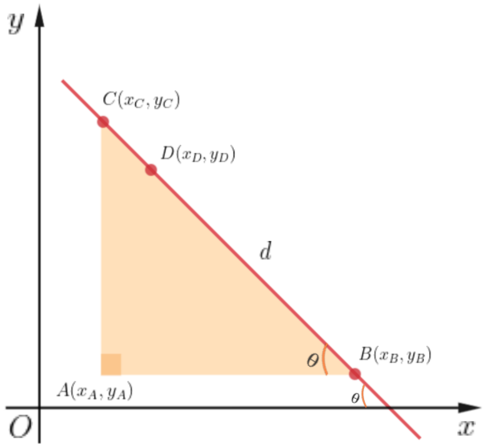


Figure 4: The slope of the line passing through points B and C

Consider $D \in \mathbb{R}^2$, a point on the hypotenuse BC with coordinates (x_D, y_D). The line which contains $D = (x_D, y_D)$ has the following equation:

$$-\frac{x_D}{\alpha} - y_D + n = 0 \Rightarrow n = \frac{x_D}{\alpha} + y_D$$

We previously computed the coordinates of point A, and because the cathetus AB and cathetus AC are parallel to the axes, we have that $y_B = y_A$ and $x_C = x_A$. Using the obtained equation of the line and the formula we found for the parameter n, we will have that:

- $-\dfrac{x_B}{\alpha} - y_A + n = 0 \Rightarrow -\dfrac{x_B}{\alpha} = y_A - n \Rightarrow -\dfrac{x_B}{\alpha} = y_A - \dfrac{x_D}{\alpha} - y_D \Rightarrow x_B = x_D + \alpha(y_D - y_A)$
- $-\dfrac{x_A}{\alpha} - y_C + n = 0 \Rightarrow y_C = -\dfrac{x_A}{\alpha} + n \Rightarrow y_C = y_D + \dfrac{x_D - x_A}{\alpha}$

As a consequence, in order to have the tighest right triangle from $\mathcal{H}_\alpha$ that covers all the points in the training set S labeled with $1(+)$, we need to find i, $i \in \{1, ..., m\}$, $y_i = 1$, such that $\forall j \in \overline{1, m}$, $y_j = 1$, we have that $-\frac{x_{i1}}{\alpha} - x_{i2} \leq -\frac{x_{j1}}{\alpha} - x_{j2}$. To compute B_S and C_S, we have:

$$a_{2S} = x_{i1} + \alpha(x_{i2} - b_{1S})$$
$$b_{2S} = x_{i2} + \frac{x_{i1} - a_{1S}}{\alpha}$$

$$B_S = (a_{2S}, b_{1S})$$
$$C_S = (a_{1S}, b_{2S})$$

If there are no positive examples in the training set (all points $x_i, i = [m]$, have label $y_i = 0$), then we choose a $Z = (z_1, z_2)$ a point that is not in the training set S and take $a_{1S} = z_1$, $b_{1S} = z_2$, $a_{2S} = a_{1S} + 2\epsilon^{-7}$ and $b_{2S} = b_{1S} + \dfrac{2\epsilon^{-7}}{\alpha}$.

The hypothesis returned by the algorithm $\mathcal{A}$, $h_S = h_{\triangle A_S B_S C_S}$ is the indicator function of the tightest rectangled triangle enclosing all positive examples, which preserves the properties of the hypotheses from H_α (see Figure 5).

Now, we want to find the sample complexity $m_H(\varepsilon, \delta)$ such that

$$\underset{S \sim \mathcal{D}^m}{\mathbb{P}}[L_{D,f}(h_S) > \varepsilon] < \delta$$

where S contains $m \geq m_H(\varepsilon, \delta)$ examples.

The area inside the right triangle $\triangle A^* B^* C^*$ is denoted by $\mathcal{T}^*$, whereas the area inside the right triangle $\triangle A_S B_S C_S$ is denoted by $\mathcal{T}_S$ (see Figure 5).

We will show that $\mathcal{T}_S \subseteq \mathcal{T}^*$.

Suppose that $T_S \nsubseteq \mathcal{T}^*$. The only case in which this can happen is if there is at least one pair $(x_k, 1) \in S$ such that $x_k = (x_{k1}, x_{k2}) \notin \mathcal{T}^*$, but we assumed realizability, thus we get a contradiction. So $\mathcal{T}_S \subseteq \mathcal{T}^*$.

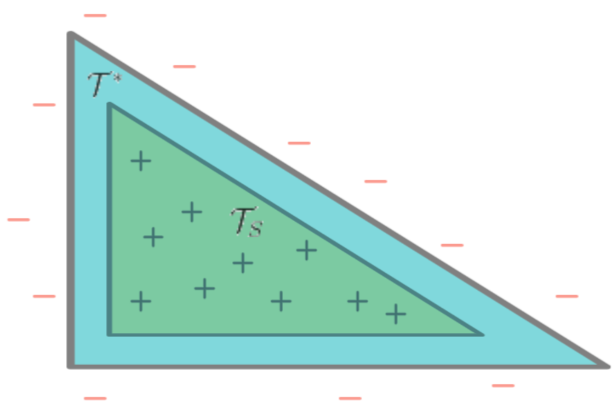


Figure 5: Triangle $\mathcal{T}_S$ is the tighest rectangle enclosing all positive examples

The definition of the algorithm $\mathcal{A}$ implies that $\mathcal{A}$ is an ERM, meaning that $L_S(h_S) = 0$, so h_S does not make any error on the training set. Considering this, we make the observation that h_S can only make errors in region $\mathcal{T}^* \backslash \mathcal{T}_S$, assigning the label $0(-)$ to points that should get label $1(+)$. All points within $\mathcal{T}_S$ will be correctly labeled (label 1), as will all points outside $\mathcal{T}^*$, which will be labeled 0.

Let's fix $\varepsilon > 0$, $\delta > 0$ and consider a distribution $\mathcal{D}$ over $\mathbb{R}^2$.

$\boxed{\textit{Case 1}}$ If $\mathcal{D}(\mathcal{T}^*) = \underset{x \sim \mathcal{D}}{\mathbb{P}}(x \in \mathcal{T}^*) \leq \varepsilon$ than in this case

$$L_{D,f}(h_S) = \underset{x \sim \mathcal{D}}{\mathbb{P}}(h_S(x) \neq f(x)) = \underset{x \sim \mathcal{D}}{\mathbb{P}}(x \in \mathcal{T}^* \backslash \mathcal{T}_S) \leq \underset{x \sim \mathcal{D}}{\mathbb{P}}(x \in \mathcal{T}^*) \leq \varepsilon$$

so we have that

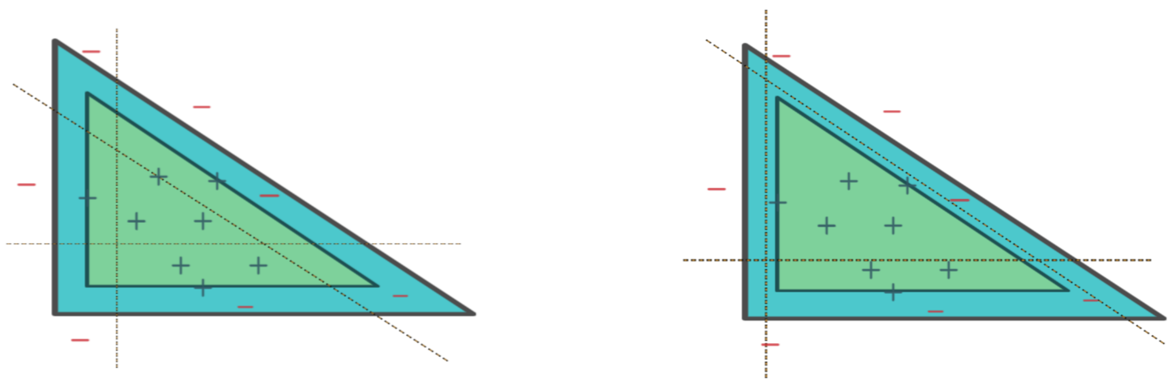
$$\underset{x \sim \mathcal{D}}{\mathbb{P}}(L_{D,f}(h_S) \leq \varepsilon) = 1 \text{ (this happens all the time).}$$

$\boxed{\textit{Case 2}}$ Let $\varepsilon > 0$ fixed and let $\mathcal{D}(\mathcal{T}^*) = \underset{x \sim \mathcal{D}}{\mathbb{P}}(x \in \mathcal{T}^*) > \varepsilon$.

Let $a_1 \in \mathbb{R}$ and $b_1 \in \mathbb{R}$ such that $a_1 \geq a_1^*$ and $b_1 \geq b_1^*$. Let d_1 denote the line which contains the point with the coordinates $(a_1, 0)$ and $d_1 \parallel O_y$. Let d_2 denote the line which contains the point with the coordinates $(0, b_1)$ and $d_2 \parallel O_x$.

We have previously determined that the line parallel to our hypotenuse has the general equation $-\frac{x}{\alpha} - y + n = 0$. Let n_1,n_1^* be numbers such that $n_1 < n_1^*$. Let d_3 denote the line with the general equation $-\frac{x}{\alpha} - y + n_1 = 0$, and the line given through the points B and C has the general equation $-\frac{x}{\alpha} - y + n_1^* = 0$.

We will define three regions $\mathcal{T}_1, \mathcal{T}_2$ and $\mathcal{T}_3$ bounded by the lines d_1, d_2, d_3 and the sides of the right triangle $\triangle A^*B^*C^*$ such that the probability masses of the trapezoids are all exactly $\frac{\varepsilon}{3}$ ($\mathcal{D}(\mathcal{T}_i) = \frac{\varepsilon}{3}, i = \overline{1,3}$), as shown in Figure 6 .



(a) $\mathcal{T}_S$ intersects each $\mathcal{T}_i$, $i = \overline{1,3}$

(b) $\mathcal{T}_S$ doesn't intersect at least one $\mathcal{T}_i$, $i = \overline{1,3}$

Figure 6: All three regions, $\mathcal{T}_1, \mathcal{T}_2$ and $\mathcal{T}_3$ have masses $\mathcal{D}(\mathcal{T}_i) = \frac{\varepsilon}{3}$, where $\varepsilon \in (0, 1)$

- If $\mathcal{T}_S$ (the region defined by the area inside the right triangle $\triangle A_S B_S C_S$ returned by $\mathcal{A}$, implemented by h_S) intersects each $\mathcal{T}_i$, $i = \overline{1,3}$:

$$L_{D,h^*}(h_S) = \underset{x\sim D}{\mathbb{P}}\left(h_S(x) \neq h^*(x)\right) = \underset{x\sim D}{\mathbb{P}}\left(x \in \mathcal{T}^* \backslash \mathcal{T}_S\right) \leq \underset{x\sim D}{\mathbb{P}}\left(x \in \bigcup_{i=1}^{3} \mathcal{T}_i\right) \leq$$

$$\leq \sum_{i=1}^{3} \underset{x\sim D}{\mathbb{P}}\left(x \in \mathcal{T}_i\right) = 3 \cdot \frac{\varepsilon}{3} = \varepsilon$$

$$\underset{x\sim D}{\mathcal{P}}\left(L_{D,f}(h_S) \leq \varepsilon\right) = 1 \quad \text{(this happens all the time)}$$

- In order to have $L_{D,h^*}(h_S) > \varepsilon$, we need that $\mathcal{T}_S$ will not intersect at least one trapezoid $\mathcal{T}_i$. We define this event with F_i, $F_i = \{S \sim D^m | \mathcal{T}_S \cap \mathcal{T}_i = \varnothing\}$. This leads to the following:

$$\underset{S\sim D^m}{\mathbb{P}}\left(L_{D,h^*}(h_S) < \varepsilon\right) \leq \underset{S\sim D^m}{\mathbb{P}}\left(\mathcal{T}_1 \cup \mathcal{T}_2 \cup \mathcal{T}_3\right) \leq \sum_{i=1}^{3} \underset{S\sim D^m}{\mathcal{P}}\left(\mathcal{T}_i\right)$$

  where the last inequality follows from the Union Bound.

  The probability that all instances do not fall in $\mathcal{T}_i$ is $\underset{S\sim D^m}{\mathbb{P}}\left(F_i\right) = \left(1 - \frac{\varepsilon}{3}\right)^m \leq e^{-\frac{\varepsilon}{3}m}$. Therefore,

$$\underset{S\sim D^m}{\mathbb{P}}\left(L_{D,h^*}(h_S) < \varepsilon\right) \leq 3e^{-\frac{\varepsilon}{3}m}$$

  Thus, it is suffices to ensure that:

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

Plugging in the assumption on m, $m \geq m_H(\varepsilon, \delta) = \left\lceil \frac{3}{\varepsilon} \cdot \log\frac{3}{\delta} \right\rceil$, we conclude our proof.

$\square$

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

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

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

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

Consider the realizability assumption.

a) Compute VCdim($\mathcal{H}$).

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.

Proof. a) Using the VC-dimension definition, we will show that $VCdim(\mathcal{H}) = 2$. Therefore, we want to prove that:

1. There $\exists C, C \subset \mathbb{R}$, where $|C| = 2$, that is shattered by $\mathcal{H}$ ($VCdim(\mathcal{H}) \geq 2$).
2. $\forall C \subset \mathbb{R}$, where $|C| = 3$, C is not shattered by $\mathcal{H}$ ($VCdim(\mathcal{H}) < 3$).

Take the set of two distinct points $C = \{a, b \mid a < b\}$. For this set we can obtain all the possible labels by choosing $h_{\theta_1,\theta_2} \in \mathcal{H}_3 \subset \mathcal{H}$ and choose θ_1 and θ_2 such that we can arrange an interval over a, b that includes neither, both, or only a or b. For example, take:

- label (0, 0): Take $\theta_1 = a - 2$ and $\theta_2 = a - 1$, then $h_{\theta_1,\theta_2}(a) = 0$ and $h_{\theta_1,\theta_2}(b) = 0$.
- label (0, 1): Take $\theta_1 = \frac{a+b}{2}$ and $\theta_2 = b + 1$, then $h_{\theta_1,\theta_2}(a) = 0$ and $h_{\theta_1,\theta_2}(b) = 1$.
- label (1, 0): Take $\theta_1 = a - 1$ and $\theta_2 = \frac{a+b}{2}$, then $h_{\theta_1,\theta_2}(a) = 1$ and $h_{\theta_1,\theta_2}(b) = 0$.
- label (1, 1): Take $\theta_1 = a - 1$ and $\theta_2 = b + 1$, then $h_{\theta_1,\theta_2}(a) = 1$ and $h_{\theta_1,\theta_2}(b) = 1$.

For any set $C, |C| = 3$, we cannot represent elements of alternative labels (for example, label (1, 0, 1)). Take an arbitrary set $C = \{a, b, c\}$ and assume without loss of generality that $a \leq b \leq c$.

- $h \in \mathcal{H}_1$: No $h \in \mathcal{H}_1$ can account for the labelling $(1, 0, 1)$, because any h_{θ_1} that assign the label 1 to a must assign the label 1 to b as well, since $\theta_1 \leq a$ and $a \leq b$. So, $VCdim(\mathcal{H}_1) < 3$.
- $h \in \mathcal{H}_2$: No $h \in \mathcal{H}_2$ can account for the labelling $(1, 0, 1)$, because any threshold that assign the label 0 to b must assign the label 0 to c as well. So, $VCdim(\mathcal{H}_2) < 3$.
- $h \in \mathcal{H}_3$: If $a \in [\theta_1, \theta_2]$ and $c \in [\theta_1, \theta_2]$, by convexity of $[\theta_1, \theta_2]$, every $b \in [a, c]$ is also in $[\theta_1, \theta_2]$. Because of this, b will also get label 1, so, any such set $\{a, b, c\}$ is not shattered by $\mathcal{H}_3$. Therefore, the VC dimension of this class representation is the largest shattered set, 2.

In conclusion, any sequence of two distinct points can be shattered by $\mathcal{H}$, but we cannot find $h \in \mathcal{H} = \mathcal{H}_1 \bigcup \mathcal{H}_2 \bigcup \mathcal{H}_3$ that labels a sequence of three distinct points with label $(+, -, +)$. So, $VCdim(\mathcal{H}) = 2$.

b) According to The Fundamental Theorem Of Statistical Learning, the Vapnik–Chervonenkis dimension characterizes the PAC learnability. Considering the result obtained in the previous subsection, $VCdim(H) = 2 < \infty$, $\mathcal{H}$ is PAC learnable.

c) Consider a training set $S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$. We are in the realizability case, so there exists $h \in \mathcal{H} = \mathcal{H}_1 \bigcup \mathcal{H}_2 \bigcup \mathcal{H}_3$ that labels the examples, $y_i = h^*(x_i)$.

Consider $\mathcal{A}$ the learning algorithm that gets the training set S and outputs $h_S = \mathcal{A}(S) =$ the tighest interval containing all the positive examples.

$h_S = h_{a_S, b_S} = \mathbb{1}_{[a_S, b_S]}$, where

$$a_S = \min_{\substack{i=\overline{1,m} \\ y_i = 1}} x_i \qquad\qquad b_S = \max_{\substack{i=\overline{1,m} \\ y_i = 1}} x_i$$

If there is no $(x_i, 1) \notin S$ (S doesn't contain positive examples), take $a_S = b_S = z$ a random point such that $(z, 0) \notin S$.

We denote by $\mathcal{R}_S = [a_S, b_S]$.

From the definition of PAC-learnability, we know that $\mathcal{H} = \mathcal{H}_1 \bigcup \mathcal{H}_2 \bigcup \mathcal{H}_3$ is $(\varepsilon, \delta)-$PAC learnable if there exists a function $m_{\mathcal{H}} : (0, 1) \times (0, 1) \to \mathbb{N}$, and there exists a learning algorithm $\mathcal{A}$ with the following property: for every $\varepsilon \in (0, 1)$, for every $\delta \in (0, 1)$, for every labeling function $f \in \mathcal{H}$ (realizability case), for every distribution $\mathcal{D}$ on $\mathbb{R}^2$, when we run the learning algorithm $\mathcal{A}$ on a training set S consisting of $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ examples sampled i.i.d. from $\mathcal{D}$ and labeled by f, the learning algorithm $\mathcal{A}$ returns a hypothesis $h_S \in \mathcal{H}$ such that, with probability at least $1 - \delta$ (over the choices of examples) the real risk of h_S is smaller than ε:

$$\underset{S \sim \mathcal{D}^m}{\mathbb{P}}[L_{D,f}(h_S) \leq \varepsilon] \geq 1 - \delta \Leftrightarrow \underset{S \sim \mathcal{D}^m}{\mathbb{P}}[L_{D,f}(h_S) > \varepsilon] < \delta$$

From the construction, h_S is an ERM, meaning that $L_S(h_S) = 0$.

Let $\varepsilon > 0, \delta > 0$ and $\mathcal{D}$ a distribution over $\mathbb{R}$. We want to find how many training examples $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ do we need such that:

$$\underset{S \sim \mathcal{D}^m}{\mathbb{P}}[L_{D,f}(h_S) > \varepsilon] < \delta$$

Case 1: If $h^* = h_{\theta_1^*} = \mathbb{1}_{[\theta_1^*, +\infty)}$ (θ_1^* is a threshold such that the hypothesis $h^* = \mathbb{1}_{[\theta_1^*, +\infty)}$ achieves $L_D(h^*) = 0$).

The generalization error of h_S will be the probability masses that falls between the θ_1^* of our hypothesis h^* and a_S and between b_S and ∞. Points in this region will be assigned the label 0, but they should get label 1.

Case 1.1: If $\mathcal{D}([a^*, +\infty]) \leq \varepsilon$ then $\underset{S \sim \mathcal{D}^m}{\mathbb{P}}[L_{D,f}(h_S) > \varepsilon] = 0$.

Case 1.2: If $\mathcal{D}([a^*, +\infty]) > \varepsilon$.

Let $\mathcal{R}_1$ be the interval $[a*, a_S]$, which has probability mass ε, and $\mathcal{R}_2$ be the interval $[b_S, \infty)$, which also has probability mass ε.

If $\mathcal{R}_S \cap \mathcal{R}_1 = \emptyset$ and $\mathcal{R}_S \cap \mathcal{R}_2 = \emptyset$ then $\underset{S \sim \mathcal{D}^m}{\mathbb{P}}[L_{D,f}(h_S) > \varepsilon] = 0$.

Else, we have that $\mathbb{P}[x_i \notin \mathcal{R}_1] \leq 1 - \varepsilon$, because $\mathcal{R}_1$ has probability mass ε. Then

$$\underset{S \sim \mathcal{D}}{\mathbb{P}}[\mathcal{R}_1] = \mathbb{P}[x_1 \notin \mathcal{R}_1 \wedge x_2 \notin \mathcal{R}_1 \wedge x_m \notin \mathcal{R}_1] = (1 - \varepsilon)^m$$

where the last inequality follows by independence of the x_i's. Also, by symmetry, we obtain $\underset{S \sim \mathcal{D}}{\mathbb{P}}[\mathcal{R}_2] = (1 - \varepsilon)^m$.

Now, we can bound the probability that $L_{D,h^*}(h_S) > \varepsilon$:

$$\underset{S \sim \mathcal{D}^m}{\mathbb{P}}[L_{D,f}(h_S) > \varepsilon] \leq \underset{S \sim \mathcal{D}}{\mathbb{P}}[\mathcal{R}_1 \cup \mathcal{R}_2] \leq \underset{S \sim \mathcal{D}}{\mathbb{P}}[\mathcal{R}_1] + \underset{S \sim \mathcal{D}}{\mathbb{P}}[\mathcal{R}_2] \leq 2(1-\varepsilon)^m \leq 2e^{-\varepsilon m} < \delta \implies m > \frac{1}{\varepsilon} log(\frac{2}{\delta})$$

Case 2: If $h* = h_{\theta_2^*} = \mathbb{1}_{(-\infty, \theta_2^*)}$ (θ_2^* is a threshold such that the hypothesis $h^* = \mathbb{1}_{(-\infty, \theta_2^*)}$ achieves $L_D(h^*) = 0$).

The generelazitation error of h_S will be the probability masses that falls between $-\infty$ and a_S, and b_S and θ_2^* of our hypothesis h^*. Points in these regions are assigned label 0, but they should get label 1.

Case 2.1: If $\mathcal{D}((-\infty, \theta_2^*)) \leq \varepsilon$ then $\underset{S \sim \mathcal{D}^m}{\mathbb{P}}[L_{D,f}(h_S) > \varepsilon] = 0$.

Case 2.2: $\mathcal{D}((-\infty, \theta_2^*)) > \varepsilon$.

Let $\mathcal{R}_1$ be the interval $(-\infty, a_S)$, which has probability mass ε, and $\mathcal{R}_2$ be the interval (b_S, θ_2^*), which also has probability mass ε.

If $\mathcal{R}_S \cap \mathcal{R}_1 = \emptyset$ and $\mathcal{R}_S \cap \mathcal{R}_2 = \emptyset$ then $\underset{S \sim \mathcal{D}^m}{\mathbb{P}}[L_{D,f}(h_S) > \varepsilon] = 0$.

Else, we have that $\mathbb{P}[x_i \notin \mathcal{R}_2] \leq 1 - \varepsilon$, because $\mathcal{R}_2$ has probability mass ε. Then

$$\underset{S \sim \mathcal{D}}{\mathbb{P}}[\mathcal{R}_1] = \mathbb{P}[x_1 \notin \mathcal{R}_1 \wedge x_2 \notin \mathcal{R}_1 \wedge x_m \notin \mathcal{R}_1] = (1 - \varepsilon)^m$$

where the last inequality follows by independence of the x_i's. Also, by symmetry, we obtain $\underset{S \sim \mathcal{D}}{\mathbb{P}}[\mathcal{R}_1] = (1 - \varepsilon)^m$.

Now, we can bound the probability that $L_{D,h^*}(h_S) > \varepsilon$:

$$\underset{S\sim\mathcal{D}^m}{\mathbb{P}}[L_{D,f}(h_S) > \varepsilon] \leq \underset{S\sim\mathcal{D}}{\mathbb{P}}[\mathcal{R}_1 \cup \mathcal{R}_2] \leq \underset{S\sim\mathcal{D}}{\mathbb{P}}[\mathcal{R}_1] + \underset{S\sim\mathcal{D}}{\mathbb{P}}[\mathcal{R}_2] \leq 2(1-\varepsilon)^m \leq 2e^{-\varepsilon m} < \delta \implies m > \frac{1}{\varepsilon} log(\frac{2}{\delta})$$

Case 3: If $h* = h_{\theta_1^*, \theta_2^*} = \mathbb{1}_{[\theta_1^*, \theta_2^*]}$, then

Case 3.1: If $\mathcal{D}([a^*, b^*]) \leq \varepsilon$ then $\underset{S\sim\mathcal{D}^m}{\mathbb{P}}[L_{D,f}(h_S) > \varepsilon] = 0$.

Case 3.2: If $\mathcal{D}([a^*, b^*]) > \varepsilon$.

Build $\mathcal{R}_1$ and $\mathcal{R}_2$, $\mathcal{R}_1 = [a^*, a]$, $\mathcal{R}_2 = [b, b^*]$ such that $\mathcal{D}(\mathcal{R}_1) = \mathcal{D}(\mathcal{R}_2) = \frac{\varepsilon}{2}$.

If $\mathcal{R}_S \cap \mathcal{R}_1 = \emptyset$ and $\mathcal{R}_S \cap \mathcal{R}_2 = \emptyset$ then $\underset{S\sim\mathcal{D}^m}{\mathbb{P}}[L_{D,f}(h_S) > \varepsilon] = 0$.

Else $\underset{S\sim\mathcal{D}^m}{\mathbb{P}}[L_{D,f}(h_S) > \varepsilon] \leq 2\left(1 - \frac{\varepsilon}{2}\right)^m \leq 2e^{-\frac{\varepsilon}{2}m} < \delta \implies m > \frac{2}{\varepsilon} log\left(\frac{2}{\delta}\right)$.

Therefore, for a training set S of size $m \geq m_{\mathcal{H}(\varepsilon,\delta)} = \max\left(\frac{1}{\varepsilon}\log\left(\frac{2}{\delta}\right), \frac{2}{\varepsilon}\log\left(\frac{2}{\delta}\right)\right)$ i.i.d. samples from $\mathcal{D}$, our learning algorithm $\mathcal{A}$ obtains a hypothesis h_S with $\underset{S\sim D^m}{\mathbb{P}}[L_{D,h^*}(h_S) < \varepsilon] \geq 1-\delta$, showing that H is (ε, δ) - PAC learnable.

$\square$