Assignment 1 - Advance Machine Learning

Lupașcu Marian – Artificial Intelligence, group #407

Problem 1. Consider $\mathcal{H} = \{h_{\theta_1}: \mathbb{R} \to \{0,1\} \mid h_{\theta_1}(x) = \mathbf{1}_{[\theta_1,\infty)}, \theta_1 \in \mathbb{R}\} \cup \{h_{\theta_2}: \mathbb{R} \to \{0,1\} \mid h_{\theta_2}(x) = \{\mathbf{1}_{(-\infty,\theta_2)}, \theta_2 \in \mathbb{R}\}$. Compute $VCdim(\mathcal{H})$.

Solution problem 1:

$$\mathcal{H} = \left\{h_{\theta_1}: \mathbb{R} \to \{0,1\} \mid h_{\theta_1}(x) = \begin{cases} 1, & x \geq \theta_1 \\ 0, & otherwise \end{cases}, \theta_1 \in \mathbb{R}\right\} \cup \left\{h_{\theta_2}: \mathbb{R} \to \{0,1\} \mid h_{\theta_2}(x) = \begin{cases} 1, & x < \theta_2 \\ 0, & otherwise \end{cases}, \theta_2 \in \mathbb{R}\right\}$$

We will prove that $\begin{matrix} VCdim(\mathcal{H}) \geq 2 & (1) \\ VCdim(\mathcal{H}) < 3 & (2) \end{matrix}$ which will lead to $VCdim(\mathcal{H}) = 2$.

In the first phase we will prove (1), that is we must find a set A such that $\mathcal{H}$ shatters A. Let $A = \{a_1, a_2 \mid a_1, a_2 \in \mathbb{R}, a_1 < a_2\}$ a set with two points on the real axis. Next, for any labeling on the points in A, we must find a classifier $h_A \in \mathcal{H}$ such that the error on set A of the classifier is 0, equivalent $L_{A(h_A)} = 0$.

For the label $\begin{cases} a_1 = 0 \\ a_2 = 0 \end{cases}$

The classifier $h_{\theta_2}$with $\theta_2 \in (-\infty, a_1]$ will be considered. E.g. $\theta_2 = a_1 - 1$.

For the label $\begin{cases} a_1 = 0 \\ a_2 = 1 \end{cases}$

The classifier $h_{\theta_1}$with $\theta_1 \in (a_1, a_2]$ will be considered. E.g. $\theta_1 = \frac{a_1+a_2}{2}$.

For the label $\begin{cases} a_1 = 1 \\ a_2 = 0 \end{cases}$

The classifier $h_{\theta_2}$with $\theta_2 \in (a_1, a_2]$ will be considered. E.g. $\theta_2 = \frac{a_1+a_2}{2}$.

For the label $\begin{cases} a_1 = 1 \\ a_2 = 1 \end{cases}$

The classifier $h_{\theta_1}$with $\theta_1 \in (-\infty, a_1]$ will be considered. E.g. $\theta_1 = a_1 - 1$.

So for any labeling of the points in A, that is $2^{|A|} = 2^2 = 4$ possible labels (tested above $\forall a_1, a_2\ with\ a_1 < a_2 \in \mathbb{R}$), we found an infinity (one actually defined) of classifiers to learn that labeling. In particular, $\forall a_1, a_2, with\ a_1 < a_2 \in \mathbb{R}$, A is obtained with 2 concrete points and the demonstration still holds. So (1) is proven $\Rightarrow VCdim(\mathcal{H}) \geq |A| = 2$ (3)

To prove (2) it is necessary that $\forall A, |A| = 3$, $\mathcal{H}$ does not shatter A. Let $A = \{a_1, a_2, a_3 \mid a_i \in \mathbb{R}, i \in \overline{1,3}\}$. Because A is a set $\Rightarrow a_1 \neq a_2 \neq a_3 \neq a_1$. Without restricting the generality we can consider $a_1 < a_2 < a_3$.

Next we will show that the label (0,1,0) cannot be learned. For this we will use Reductio ad Absurdum Method: Suppose that:

$$\exists h_P \in \mathcal{H} \text{ such that } L_{A(h_A)} = 0 \Rightarrow \begin{cases} h_P \in \overbrace{\left\{ h_{\theta_1} : \mathbb{R} \to \{0,1\} \mid h_{\theta_1}(x) = \begin{cases} 1, & x \geq \theta_1 \\ 0, otherwise \end{cases}, \theta_1 \in \mathbb{R} \right\}}^{\mathcal{H}_1} \\ \qquad\qquad\qquad\qquad or \\ h_P \in \underbrace{\left\{ h_{\theta_2} : \mathbb{R} \to \{0,1\} \mid h_{\theta_2}(x) = \begin{cases} 1, & x < \theta_2 \\ 0, otherwise \end{cases}, \theta_2 \in \mathbb{R} \right\}}_{\mathcal{H}_2} \end{cases}$$

$$Case1: if\ h_P \in \mathcal{H}_1 \Rightarrow \exists \theta_1\ such\ that \begin{cases} h_P(a_1) = 0 \\ h_P(a_2) = 1 \\ h_P(a_3) = 0 \end{cases}, where\ h_P = \begin{cases} 1, & x \in [\theta_1, \infty) \\ 0, & otherwise \end{cases} \Rightarrow$$

$$\left. \begin{cases} \begin{cases} \mathrm{a}_2 \in [\theta_1, \infty) \\ \{a_1, a_3\} \subset (-\infty, \theta_1) \end{cases} \Rightarrow \begin{cases} a_2 \geq \theta_1 \\ a_1 < \theta_1 \\ a_3 < \theta_1 \end{cases} \\ a_1 < a_2 < a_3 \end{cases} \right| \Rightarrow \begin{cases} a_2 \geq \theta_1 \\ a_1 < a_3 < \theta_1 \end{cases} \Rightarrow \begin{cases} a_1 < a_3 < \theta_1 \leq a_2 \\ a_1 < a_2 < a_3 \end{cases} \Rightarrow \perp$$

$$\left. \begin{array}{l} \Rightarrow \nexists h_P \in \mathcal{H}_1 \text{ such that } L_{A(h_A)} = 0 \\ Analogous\ Case2\text{: } \nexists h_P \in \mathcal{H}_2 \text{ such that } L_{A(h_A)} = 0 \\ \mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2 \end{array} \right| \Rightarrow \nexists h_P \in \mathcal{H} \text{ such that } L_{A(h_A)} = 0$$

$$\Rightarrow VCdim(\mathcal{H}) < |\mathrm{A}| = 3 \ (4)$$

$$\left. \begin{array}{r} (3) \\ (4) \end{array} \right| \Rightarrow \boxed{VCdim(\mathcal{H}) = 2} \ \blacksquare$$

Problem 2. Consider $\mathcal{H}$ to be the class of all centered in origin sphere classifiers in the 3D space. A centered in origin sphere classifier in the 3D space is a classifier h_r that assigns the value 1 to a point if and only if it is inside the sphere with radius $r > 0$ and center given by the origin $\boldsymbol{O}(0,0,0)$. Consider the realizability assumption.

a. show that the class $\mathcal{H}$ can be (ε, δ) − PAC learned by giving the algorithm A and determining the sample complexity $m_{\mathcal{H}}(\varepsilon, \delta)$ such that the definition of PAC − learnability is satisfied.
b. compute $VCdim(\mathcal{H})$.

Solution problem 2 a:

$$\mathcal{H} = \left\{ h_{\mathrm{r}}: \mathbb{R}^3 \to \{0,1\} \mid h_{\mathrm{r}}\big(x = (a,b,c)\big) = \begin{cases} 1, & a^2 + b^2 + c^2 \le r \\ 0, & otherwise \end{cases}, \mathrm{r} \in \mathbb{R}_+^* \right\}$$

How are we in realizability case $\Rightarrow \exists h^* \in \mathcal{H}$ such that $L_{h^*,D}(h^*) = 0$ (RA). $h^* \stackrel{\text{def}}{=} h_{r^*} = S(O(0,0,0), r^*) \stackrel{\text{def}}{=} S(O, r^*)\ with\ r^* > 0$

$\mathcal{H}$ is PAC − learnable if $\begin{cases} \exists m_{\mathcal{H}}: (0,1)^2 \to \mathbb{N} \\ \exists\ A - learning\ algorithm\ with\ the\ following\ property \end{cases}$

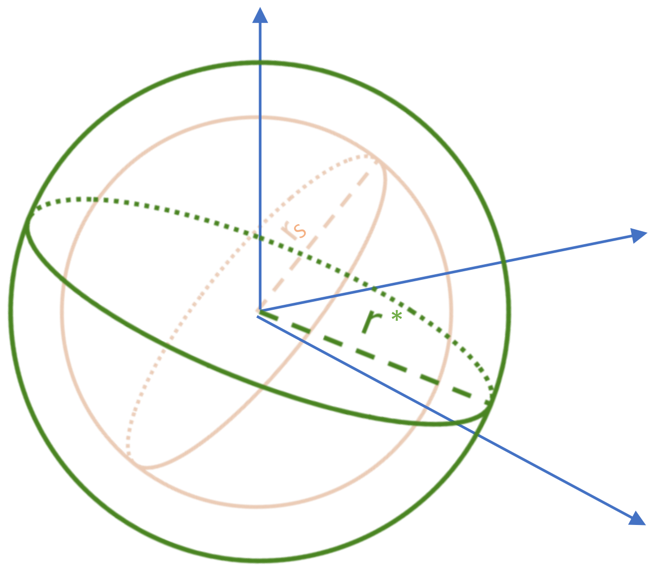
$$\forall \varepsilon \forall \delta \forall f \forall D\ we\ have\ that\ \underset{S\sim D^m}{\mathrm{P}}\left[L_{h^*,D}(h_S) \le \varepsilon\right] \ge \delta \Leftrightarrow \underset{S\sim D^m}{\mathrm{P}}\left[L_{h^*,D}(h_S) > \varepsilon\right] < \delta\ where:$$

- $\varepsilon, \delta \in (0,1)$
- $f \in \mathcal{H}$ labeling function
- D distribution on $\mathbb{R}^3$
- S training set, $|S| = m < m_{\mathcal{H}}(\varepsilon, \delta)$ examples sampled i.i.d. from D and labeled by f
- $h_S = A(S)$ with $L_S(h_S) = 0$

First we need to find the algorithm A.

Notation: Let $\|X\| = |OX|, where\ X \in \mathbb{R}^3, X = (a,b,c), O = (0,0,0)$ ($\|X\|$distance from X to origin).

Let $S = \{(x_i, y_i) \mid x_i \in \mathbb{R}^3, y_i \in \{0,1\}, y_i = h_{r^*}(x_i), x_i = (a_i, b_i, c_i), i \in \overline{1,m}\}$ the training set.



Consider the following algorithm A, which has as input the training set S and returns the hypothesis h_S:

$$\mathrm{A(S)} = h_S \stackrel{\text{def}}{=} h_{r_S}\ \text{where}\ r_S = \max_{\substack{i\in 1,m \\ y_i=1 \\ x_i=(a_i,b_i,c_i)}} a_i^2 + b_i^2 + c_i^2 = \max_{\substack{i\in 1,m \\ y_i=1}} \|x_i\|$$

That is, the algorithm returns a sphere of minimum size encompassing all points with label 1 in S. If there are no such points in S ($\nexists i = \overline{1,m}\ such\ that\ y_i = 1$) then we do not have positive learning examples and it is natural to assign label 0 to all points in this sense (sphere centered in origin of radius 0) $\Leftrightarrow \nexists i = \overline{1,m}\ such\ that\ y_i = 1 \Rightarrow \mathrm{A(S)} = h_S = h_0 = S(O(0,0,0), 0)$.

Furthermore,

$$r_S = \max_{\substack{i=\in \\ y_i=1}} a^2 + b^2 + c^2 \leq r^* \Rightarrow L_S(h_S) = \frac{|\{i \mid i \in \overline{1,m}, h_S(x_i) \neq y_i, (x_i, y_i) \in S\}|}{m}$$

$$= \frac{|\{i \mid i \in \overline{1,m}, h_S(x_i) = 0, y_i = 1\,, (x_i, y_i) \in S\}| + |\{i \mid i \in \overline{1,m}, h_S(x_i) = 1, y_i = 0\,, (x_i, y_i) \in S\}|}{m}$$

$$= \frac{|\{i \mid i \in \overline{1,m}, \|x_i\| > r_S, y_i = 1\,, (x_i, y_i) \in S\}| + |\{i \mid i \in \overline{1,m}, \|x_i\| \leq r_S, y_i = 0\,, (x_i, y_i) \in S\}|}{m}$$

$$= \frac{|\{i \mid i \in \overline{1,m}, \|x_i\| > r_S, \|x_i\| \leq r^*\,, (x_i, y_i) \in S\}| + |\{i \mid i \in \overline{1,m}, \|x_i\| \leq r_S, \|x_i\| > r^*\,, (x_i, y_i) \in S\}|}{m}$$

$$= \frac{|\{i \mid i \in \overline{1,m}, \|x_i\| \in (r_S, r^*]\,, (x_i, y_i) \in S\}| + |\emptyset|}{m} = \frac{|\emptyset| + |\emptyset|}{m} = 0 \Rightarrow A \text{ is ERM}$$

$$\textit{We assume that } \{i \mid i \in \overline{1,m}, \|x_i\| \in (r_S, r^*]\,, (x_i, y_i) \in S\} \neq \emptyset \Rightarrow \exists j \in \overline{1,m}, (x_j, y_j) \in S \ \ s.t. \|x_j\| \in (r_S, r^*] \Rightarrow \begin{cases} \|x_j\| > r_S \Rightarrow y_j = 0 \ (\textit{otherwise the algorithm would have set } r_S \textit{ with minimum } \|x_j\|) \\ \|x_j\| \leq r^* \Rightarrow y_j = 1 \left(\textit{because } x_j \in S(O, r^*)\right) \end{cases}$$

$$\Rightarrow \begin{cases} y_j = 0 \\ y_j = 1 \end{cases} \Rightarrow \perp \Rightarrow \{i \mid i \in \overline{1,m}, \|x_i\| \in (r_S, r^*]\,, (x_i, y_i) \in S\} = \emptyset$$

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

$$\mathop{\mathrm{P}}_{S\sim D^m}\left[L_{h^*,D}(h_S) \leq \varepsilon\right] \geq \delta \Leftrightarrow \mathop{\mathrm{P}}_{S\sim D^m}\left[L_{h^*,D}(h_S) > \varepsilon\right] < \delta$$

where S training set, $|S| = m < \ m_{\mathcal{H}}(\varepsilon, \delta)$ examples sampled i.i.d. from D.

WE can see 3 areas of interest

The only error-causing area is $S(O, r^*) \setminus S(O, r_S)$ because here are the only points that have the label 1 but are labeled with 0 by h_S. In addition the points in $\mathbb{R}^3 / S(O, r^*)$ have label 0 and are labeled with 0 by h_S ✓, and the points in $S(O, r_S)$ have label 1 and are labeled with 1 by h_S ✓.

Let's fixe $\varepsilon \in (0,1)$ *and* $\delta \in (0,1)$ and let D distribution over $\mathbb{R}^3$. Here are two cases:

Case 1: $D\left(S(O, r^*)\right) = \mathop{\mathrm{P}}_{x\sim D}[x \in S(O, r^*)] \leq \varepsilon$, in this case:

$$L_{h^*,D}(h_S) = \mathop{\mathrm{P}}_{x\sim D}[h_S(x) \neq h^*(x)] = \mathop{\mathrm{P}}_{x\sim D}[x \in S(O, r^*)/S(O, r_S)] \leq \mathop{\mathrm{P}}_{x\sim D}[x \in S(O, r^*)] \leq \varepsilon \Rightarrow \mathop{\mathrm{P}}_{S\sim D^m}\left[L_{h^*,D}(h_S) \leq \varepsilon\right] = 1 > \delta \in (0,1) \ (1)$$

Case 2: $D\left(S(O, r^*)\right) = \mathop{\mathrm{P}}_{x\sim D}[x \in S(O, r^*)] > \varepsilon$.

Let $S(O, k)$ *such that* $D\left(S(O, r^*)/S(O, k)\right) = \varepsilon \Leftrightarrow \mathop{\mathrm{P}}_{x\sim D}[x \in S(O, r^*)/S(O, k)] = \varepsilon$

Case 2.1: $S(O, r_S) \cap \left(S(O, r^*)/S(O, k)\right) \neq \emptyset \Leftrightarrow k < r_S$

$$L_{h^*,D}(h_S) = \mathop{\mathrm{P}}_{x\sim D}\left[h_{S(x)} \neq h^*(x)\right] = \mathop{\mathrm{P}}_{x\sim D}[x \in S(O, r^*)/S(O, r_S)] \leq \mathop{\mathrm{P}}_{x\sim D}[x \in S(O, r^*)/S(O, k)] = \varepsilon \Rightarrow \mathop{\mathrm{P}}_{S\sim D^m}\left[L_{h^*,D}(h_S) \leq \varepsilon\right] = 1 > \delta \in (0,1) \ (2)$$

Case 2.1: $S(O, r_S) \cap \left(S(O, r^*)/S(O, k)\right) = \emptyset \Leftrightarrow r_S < k$ then we have to enlarge S.

Let $F = \left\{S\sim D^m \mid S(O, r_S) \cap \left(S(O, r^*)/S(O, k)\right) = \emptyset\right\}$

$$\underset{S\sim D^m}{\mathrm{P}}\left[L_{h^*,D}(h_S) > \varepsilon\right] = \underset{S\sim D^m}{\mathrm{P}}[F]$$

$$= the\ probability\ that\ S(O,r_S)\ will\ not\ intersect\ \underbrace{S(O,r^*)/S(O,k)}_{\varepsilon}\ (S\ is\ i.i.d.)$$

$$= (1-\varepsilon)^m(1-x \le e^{-x}) \le e^{-\varepsilon m} < \delta \Rightarrow m > \frac{\ln\frac{1}{\delta}}{\varepsilon}$$

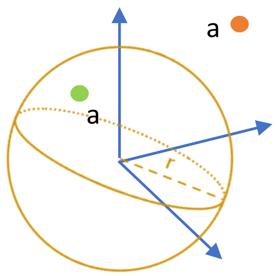
$$\boxed{m \ge m_{\mathcal{H}}(\varepsilon,\delta) = \left\lceil \frac{\ln\frac{1}{\delta}}{\varepsilon} \right\rceil} (3)$$

$$(1),(2),(3) \Rightarrow \mathcal{H} \text{ can be } (\varepsilon,\delta) - \text{PAC learned by A (defined above)} \blacksquare$$

Solution problem 2 b:

We will prove that $\begin{matrix} VCdim(\mathcal{H})\ge 1 & (1) \\ VCdim(\mathcal{H})<2 & (2) \end{matrix}$ which will lead to $VCdim(\mathcal{H}) = 1$.

In the first phase we will prove (1), that is we must find a set A such that $\mathcal{H}$ shatters A. Let $A = \{a|\ a \in \mathbb{R}^3\}$ a set with only one point in space. Next, for any labeling on the points in A, we must find a classifier $h_A \in \mathcal{H}$ such that the error on set A of the classifier is 0, equivalent $L_{A(h_A)} = 0$.



For the label $a = 0$

The classifier h_{r_A} with $r_A \in (\|a\|, \infty)$ will be considered. E.g. $r_A = \|a\| + 1$.

For the label $a = 1$

The classifier h_{r_A} with $r_A \in (0, \|a\|\,]$ will be considered. E.g. $r_A = \frac{\|a\|}{2}$.

So for any labeling of the points in A, that is $2^{|A|} = 2^1 = 2$ possible labels (tested above $\forall a \in \mathbb{R}^3$), we found infinity (one actually defined) of classifiers to learn that labeling. In particular, $\forall a \in \mathbb{R}^3$, A is obtained with 1 concrete point and the demonstration still holds. So (1) is proven $\Rightarrow VCdim(\mathcal{H}) \ge |A| = 1$ (3)

To prove (2) it is necessary that $\forall A, |A| = 2,\ \mathcal{H}$ does not shatter A. Let $A = \{a_1, a_2|\ a_i \in \mathbb{R}^3, i \in \overline{1,2}\}$. There are 2 cases.

Case 1: $\|a_1\| \ne \|a_2\|$. Without restricting the generality we can consider $\|a_1\| < \|a_2\|$.

Next we will show that the label (0,1) cannot be learned. For this we will use Reductio ad Absurdum Method: Suppose that:

$$\exists h_P \in \mathcal{H} \text{ such that } L_{A(h_A)} = 0 \Rightarrow \exists r_P\ such\ that \begin{cases} h_P(a_1) = 0 \\ h_P(a_2) = 1 \end{cases} \Rightarrow \begin{cases} \|a_1\| > r_P \\ \|a_2\| \le r_P \end{cases} \left. \begin{array}{r} \Rightarrow \|a_2\| \le r_P < \|a_1\| \\ But\ \|a_1\| < \|a_2\| \end{array} \right|$$

$$\Rightarrow \perp \Rightarrow \nexists h_P \in \mathcal{H} \text{ such that } L_{A(h_A)} = 0 \Rightarrow VCdim(\mathcal{H}) < |A| = 2\ (4)$$

Case 2: $\|a_1\| = \|a_2\|$

Next we will show that the label (0,1) cannot be learned. For this we will use Reductio ad Absurdum Method: Suppose that:

$$\exists h_P \in \mathcal{H} \text{ such that } L_{A(h_A)} = 0 \Rightarrow \exists r_P\ such\ that \begin{cases} h_P(a_1) = 0 \\ h_P(a_2) = 1 \end{cases} \Rightarrow \begin{cases} \|a_1\| > r_P \\ \|a_2\| \le r_P \end{cases} \left. \begin{array}{r} \Rightarrow \|a_2\| \le r_P < \|a_1\| \\ But\ \|a_1\| = \|a_2\| \end{array} \right|$$

$$\Rightarrow \perp \Rightarrow \nexists h_P \in \mathcal{H} \text{ such that } L_{A(h_A)} = 0 \Rightarrow VCdim(\mathcal{H}) < |A| = 2\ (5)$$

$$(3),(4),(5) \Rightarrow \boxed{VCdim(\mathcal{H}) = 1} \blacksquare$$

Problem 3. Let $\mathcal{H} = \left\{h_\theta : \mathbb{R} \to \{0,1\} \mid h_\theta(x) = \mathbf{1}_{[\theta,\theta+1]\cup[\theta+2,\infty)}, \theta \in \mathbb{R}\right\}$. Compute $VCdim(\mathcal{H})$.

Solution problem 3:

$$\mathcal{H} = \left\{h_\theta : \mathbb{R} \to \{0,1\} \mid h_\theta(x) = \begin{cases} 1, & x \in [\theta, \theta+1] \cup [\theta+2, \infty) \\ 0, & otherwise \end{cases}, \theta \in \mathbb{R}\right\}$$

We will prove that $\begin{matrix} VCdim(\mathcal{H}) \geq 3 & (1) \\ VCdim(\mathcal{H}) < 4 & (2) \end{matrix}$ which will lead to $VCdim(\mathcal{H}) = 3$.

In the first phase we will prove (1), that is we must find a set A such that $\mathcal{H}$ shatters A. Let $A = \{a - 0.75, a, a + 0.75 \mid a \in \mathbb{R}\}$ a set with three points on the real axis. Next, for any labeling on the points in A, we must find a classifier $h_A \in \mathcal{H}$ such that the error on set A of the classifier is 0, equivalent $L_{A(h_A)} = 0$.

For the label $\begin{cases} a_1 = 0 \\ a_2 = 0. \\ a_3 = 0 \end{cases}$

The classifier h_θ with $\theta \in (a + 0.75, \infty)$ will be considered. E.g. $\theta = a + 1$.

For the label $\begin{cases} a_1 = 0 \\ a_2 = 0. \\ a_3 = 1 \end{cases}$

The classifier h_θ with $\theta \in (a, a + 0.75]$ will be considered. E.g. $\theta = a + 0.5$.

For the label $\begin{cases} a_1 = 0 \\ a_2 = 1. \\ a_3 = 0 \end{cases}$

The classifier h_θ with $\theta \in (a - 0.75, a - 0.25)$ will be considered. E.g. $\theta = a - 0.5$.

For the label $\begin{cases} a_1 = 1 \\ a_2 = 0. \\ a_3 = 0 \end{cases}$

The classifier h_θ with $\theta \in (a - 1.25, a - 1)$ will be considered. E.g. $\theta = a - 1.1$

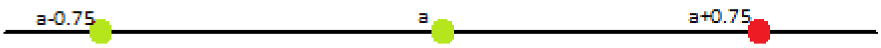
For the label $\begin{cases} a_1 = 0 \\ a_2 = 1. \\ a_3 = 1 \end{cases}$

The classifier h_θ with $\theta \in [a - 2, a - 1.75)$ will be considered. E.g. $\theta = a - 2$.

For the label $\begin{cases} a_1 = 1 \\ a_2 = 0. \\ a_3 = 0 \end{cases}$

The classifier h_θ with $\theta \in [a - 1.75, a - 1.25]$ will be considered. E.g. $\theta = a - 1.75$.

For the label $\begin{cases} a_1 = 1 \\ a_2 = 1. \\ a_3 = 0 \end{cases}$



The classifier h_θ with $\theta \in [a - 1, a - 0.75]$ will be considered. E.g. $\theta = a - 1$.

For the label $\begin{cases} a_1 = 1 \\ a_2 = 1. \\ a_3 = 1 \end{cases}$

The classifier h_θ with $\theta \in (-\infty, a - 2.75]$ will be considered. E.g. $\theta = a - 3$.

So for any labeling of the points in A, that is $2^{|A|} = 2^3 = 8$ possible labels (tested above $\forall a \in \mathbb{R}$), we found an infinity (one actually defined) of classifiers to learn that labeling. In particular, for example $a = 0$, A is obtained with 2 concrete points ($\{-0.75, 0, 0.75\}$) and the demonstration still holds. So (1) is proven $\Rightarrow VCdim(\mathcal{H}) \geq |\text{A}| = 3$ (3)

To prove (2) it is necessary that $\forall A, |A| = 4$, $\mathcal{H}$ does not shatter A. Let $A = \{a_1, a_2, a_3, a_4 |\ a_i \in \mathbb{R}, \text{i} \in \overline{1,4}\}$. Because A is a set $\Rightarrow a_i \neq a_j, \forall i \neq j \in \overline{1,4}$. Without restricting the generality we can consider $a_1 < a_2 < a_3 < a_4$.

Next we will show that the label (0,1,0,1) cannot be learned. For this we will use Reductio ad Absurdum Method: Suppose that:

$$\exists h_P \in \mathcal{H} \text{ such that } L_{A(h_A)} = 0 \Rightarrow \exists \theta_P \ such\ that \begin{cases} h_P(a_1) = 1 \\ h_P(a_2) = 0 \\ h_P(a_3) = 1 \\ h_P(a_4) = 0 \end{cases} \Rightarrow \begin{cases} a_1 \in [\theta_P, \theta_P + 1] or [\theta_P + 2, \infty) & (4) \\ a_2 \in (-\infty, \theta_P) or (\theta_P + 1, \theta_P + 2) & (5) \\ a_3 \in [\theta_P, \theta_P + 1] or [\theta_P + 2, \infty) & (6) \\ a_4 \in (-\infty, \theta_P) or (\theta_P + 1, \theta_P + 2) & (7) \end{cases}$$

$$(7) \Rightarrow \left.\begin{cases} a_4 \in (-\infty, \theta_P) \\ or \\ a_4 \in (\theta_P + 1, \theta_P + 2) \end{cases}\right| \Rightarrow \left.\begin{matrix} a_4 < \theta_P + 2 \\ \\ a_1 < a_2 < a_3 < a_4 \end{matrix}\right| \Rightarrow \left.\begin{matrix} a_1 < \theta_P + 2 \\ \\ (4) \end{matrix}\right| \Rightarrow a_1 \in [\theta_P, \theta_P + 1]\ (8)$$

$$\left.\begin{matrix} (8) \Rightarrow a_1 \geq \theta_P \\ a_1 < a_2 < a_3 < a_4 \end{matrix}\right| \Rightarrow \left.\begin{matrix} a_2 \geq \theta_P \\ (5) \end{matrix}\right| \Rightarrow a_2 \in (\theta_P + 1, \theta_P + 2)\ \ (9)$$

$$\left.\begin{matrix} (9) \Rightarrow a_2 > \theta_P + 1 \\ a_1 < a_2 < a_3 < a_4 \end{matrix}\right| \Rightarrow \left.\begin{matrix} a_3 > \theta_P + 1 \\ (6) \end{matrix}\right| \Rightarrow a_3 \in [\theta_P + 2, \infty)\ \ (10)$$

$$\left.\begin{matrix} (10) \Rightarrow a_3 \geq \theta_P + 2 \\ a_1 < a_2 < a_3 < a_4 \end{matrix}\right| \Rightarrow \left.\begin{matrix} a_4 \geq \theta_P + 2 \\ (7) \end{matrix}\right| \Rightarrow \perp \Rightarrow \nexists h_P \in \mathcal{H} \text{ such that } L_{A(h_A)} = 0 \Rightarrow VCdim(\mathcal{H}) < |\text{A}| = 4\ (11)$$

$$(3), (11) \Rightarrow \boxed{VCdim(\mathcal{H}) = 3} \blacksquare$$

Problem 4. An axis aligned square classifier in the plane is a classifier that assigns the value 1 to a point if and only if it is inside a certain square. Formally, given the real numbers $a_1, a_2, r > 0 \in \mathbb{R}$ we define the classifier $h_{(a_1,a_2,r)}$ by

$$h_{(a_1,a_2,r)}(x_1,x_2) = \begin{cases} 1, if\ a_1 \le x_1 \le a_1 + r, a_2 \le x_2 \le a_2 + r \\ 0, otherwise \end{cases}$$

The class of all axis aligned squares in the plane is defined as:

$$\mathcal{H} = \{h_{(a_1,a_2,r)}: \mathbb{R}^2 \to \{0,1\} \mid a_1, a_2, r \in \mathbb{R}, r > 0\}$$

Consider the realizability assumption.

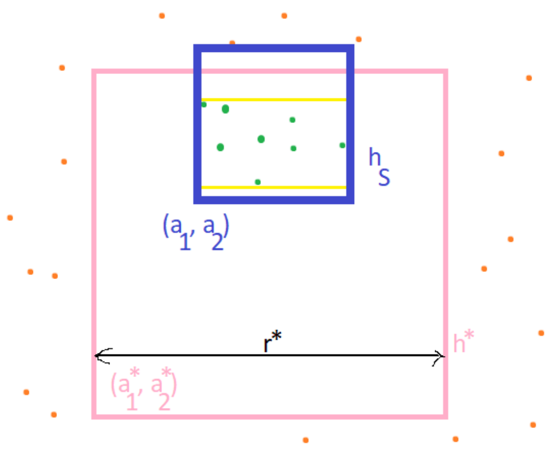
a. give a learning algorithm A that return a hypothesis h_S from $\mathcal{H}$, $h_S = A(S)$ consistent with the training set S (h_S has empirical risk 0 on S);
b. find the sample complexity $m_{\mathcal{H}}(\varepsilon, \delta)$ in order to show that $\mathcal{H}$ is PAC – learnable;
c. compute $VCdim(\mathcal{H})$.

How are we in realizability case $\Rightarrow \exists h^* \in \mathcal{H}$ such that $L_{h^*,D}(h^*) = 0$ (RA). $h^* \overset{\text{def}}{=} h_{(a_1^*,a_2^*,r^*)} =$ $with\ a_1^*, a_2^*, r^* \in \mathbb{R}, r^* > 0$

Solution problem 4 a:

Let $S = \{(x_i, y_i) \mid x_i \in \mathbb{R}^2, y_i \in \{0,1\}, y_i = h^*(x_i), x_i = (a_i, b_i), i \in \overline{1,m}\}$ the training set.

Consider the following algorithm A, which has as input the training set S and returns the hypothesis $h_{(a_1,a_2,r)}$:



1. $$\begin{cases} l = \min\limits_{\substack{i \in \overline{1,m} \\ y_i = 1 \\ x_i = (a_i,b_i)}} a_i \\ r = \max\limits_{\substack{i \in \overline{1,m} \\ y_i = 1 \\ x_i = (a_i,b_i)}} a_i \\ u = \max\limits_{\substack{i \in \overline{1,m} \\ y_i = 1 \\ x_i = (a_i,b_i)}} b_i \\ d = \min\limits_{\substack{i \in \overline{1,m} \\ y_i = 1 \\ x_i = (a_i,b_i)}} b_i \end{cases} \quad \begin{cases} W = r - l \\ H = u - d \end{cases}$$

2. $$\begin{cases} \begin{cases} a_1 = l \\ a_2 = d \\ r = W = H \end{cases} & , if\ W = H \\ \begin{cases} \begin{cases} a_1 = l \\ a_2 = d \\ r = W \end{cases} & , if\ m = \infty \\ \begin{cases} a_1 = l \\ a_2 = d - (m - u), otherwise \\ r = W \end{cases} \end{cases} & , if\ W > H, where\ m = \min\left(\min\limits_{\substack{i \in \overline{1,m} \\ y_i = 0 \\ x_i = (a_i,b_i) \\ l \le a_i \le r}} b_i, \infty \right) \\ \begin{cases} \begin{cases} a_1 = l \\ a_2 = d\ , if\ m = \infty \\ r = H \end{cases} \\ \begin{cases} a_1 = l - (m - r) \\ a_2 = d, otherwise \\ r = H \end{cases} \end{cases} & , if\ W < H, where\ m = \min\left(\min\limits_{\substack{i \in \overline{1,m} \\ y_i = 0 \\ x_i = (a_i,b_i) \\ d \le b_i \le u}} a_i, \infty \right) \end{cases}$$

3. Return $h_{(a_1,a_2,r)}$

In the first 1 the smallest rectangle that includes all the points with the label 1 is determined (l u r d) and its width and height are calculated. In step 2 the parameters of the classifier are calculated. If the previous rectangle is square, then it is returned. If it is a rectangle with a length greater than the width, then the rectangle extends upwards as much as possible (up to a label 0) if it can no longer extend downwards. Analogous for rectangles with a width greater than length.

If there are no such points in S ($\nexists i = \overline{1,m}$ *such that* $y_i = 1$) then we do not have positive learning examples and it is natural to assign label 0 to all points in this sense $\Leftrightarrow \nexists i = \overline{1,m}$ *such that* $y_i = 1$ $\Rightarrow$ A(S) $= h_S = h_{P,0}$, P a random point in $\mathbb{R}^2$.

$$L_S(h_S) = \frac{|\{i \mid i \in \overline{1,m}, h_S(x_i) \neq y_i, (x_i, y_i) \in S\}|}{m}$$

$if\ W = H$:

$$L_S(h_S) = \\ = \frac{|\{i \mid i \in \overline{1,m}, h_S(x_i) = 0, y_i = 1, (x_i, y_i) \in S\}| + |\{i \mid i \in \overline{1,m}, h_S(x_i) = 1, y_i = 0, (x_i, y_i) \in S\}|}{m} = \\ = \frac{|\{i \mid i \in \overline{1,m}, (x_i, y_i) \in [a_1^*, a_1^* + r^*] \times [a_2^*, a_2^* + r^*] \backslash [a_1, a_1 + r] \times [a_2, a_2 + r], (x_i, y_i) \in S\}| + |\emptyset|}{m} \\ = \frac{|\emptyset| + |\emptyset|}{m} = 0 \ (\nexists (x_i, y_i) \in [a_1^*, a_1^* + r^*] \times [a_2^*, a_2^* + r^*] \backslash [a_1, a_1 + r] \times [a_2, a_2 + r] \ni \mathrm{S})$$

if point $(x_i, y_i) \ni S$ *then the algorithm would update the parameters so that* $(x_i, y_i) \epsilon [a_1, a_1 + r] \times [a_2, a_2 + r]$ (1)

$if\ W > H$:

$$L_S(h_S) = \\ = \frac{|\{i \mid i \in \overline{1,m}, h_S(x_i) = 0, y_i = 1, (x_i, y_i) \in S\}| + |\{i \mid i \in \overline{1,m}, h_S(x_i) = 1, y_i = 0, (x_i, y_i) \in S\}|}{m} = \\ = \frac{|\{i \mid i \in \overline{1,m}, (x_i, y_i) \in [a_1^*, a_1^* + r^*] \times [a_2^*, a_2^* + r^*] \backslash [a_1, a_1 + r] \times [a_2, a_2 + r], (x_i, y_i) \in S\}| + |\{i \mid i \in \overline{1,m}, (x_i, y_i) \in [a_1, a_1 + r] \times [a_2, a_2 + r] \backslash [a_1^*, a_1^* + r^*] \times [a_2^*, a_2^* + r^*], (x_i, y_i) \in S\}|}{m} \\ = \frac{|\emptyset| + |\emptyset|}{m} = 0 \ \ (2)$$

From the construction of the algorithm all the points with label 1 in S are included in the rectangle of step 1, a rectangle that is strictly included in the square returned by the algorithm$\Rightarrow \{i \mid i \in \overline{1,m}, (x_i, y_i) \in [a_1^*, a_1^* + r^*] \times [a_2^*, a_2^* + r^*] \backslash [a_1, a_1 + r] \times [a_2, a_2 + r], (x_i, y_i) \in S\} = \emptyset$ (*false negative*).

From the construction of the algorithm the rectangle from step 1, extends upwards to the first point with the label 0, so no point with the label 0 is included in the square returned by the algorithm $\Rightarrow \{i \mid i \in \overline{1,m}, (x_i, y_i) \in [a_1, a_1 + r] \times [a_2, a_2 + r] \backslash [a_1^*, a_1^* + r^*] \times [a_2^*, a_2^* + r^*], (x_i, y_i) \in S\} = \emptyset$ (*false positive*).

if $W < H$ *analogous case* $W > H(3)$

$$\boxed{(1), (2), (3) \Rightarrow A \text{ is ERM}}$$

Solution problem 4 b:

$\mathcal{H}$ is PAC – learnable if $\begin{cases} \exists m_{\mathcal{H}}: (0,1)^2 \to \mathbb{N} \\ \exists\ A - learning\ algorithm\ with\ the\ following\ property \end{cases}$

$$\forall \varepsilon \forall \delta \forall f \forall D \text{ we have that } \underset{S \sim D^m}{\mathrm{P}}\left[L_{h^*,D}(h_S) \leq \varepsilon\right] \geq \delta \Leftrightarrow \underset{S \sim D^m}{\mathrm{P}}\left[L_{h^*,D}(h_S) > \varepsilon\right] < \delta \text{ where:}$$

- $\varepsilon, \delta \in (0,1)$

- $f \in \mathcal{H}$ labeling function
- D distribution on $\mathbb{R}^3$
- S training set, $|S| = m < m_{\mathcal{H}}(\varepsilon, \delta)$ examples sampled i.i.d. from D and labeled by f
- $h_S = A(S)$ with $L_S(h_S) = 0$

Notation: $\begin{cases} [a_1^*, a_1^* + r^*] \times [a_2^*, a_2^* + r^*] \stackrel{\text{def}}{=} P^* \\ [a_1, a_1 + r] \times [a_2, a_2 + r] \stackrel{\text{def}}{=} P_S \end{cases}$

First we need to find the algorithm A, Let's take A from 4.a.

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

$$\underset{S\sim D^m}{\mathrm{P}}\left[L_{h^*,D}(h_S) \le \varepsilon\right] \ge \delta \Leftrightarrow \underset{S\sim D^m}{\mathrm{P}}\left[L_{h^*,D}(h_S) > \varepsilon\right] < \delta$$

where S training set, $|S| = m < m_{\mathcal{H}}(\varepsilon, \delta)$ examples sampled i.i.d. from D.

$if\ W = H$:

We can see 3 areas of interest:

The only error-causing area is P^*/P_S because here are the only points that have the label 1 but are labeled with 0 by h_S. In addition the points in $\mathbb{R}^3/P^*$ have label 0 and are labeled with 0 by h_S ✓, and the points in P_S have label 1 and are labeled with 1 by h_S ✓.

Let's fixe $\varepsilon \in (0,1)\ and\ \delta \in (0,1)$ and let D distribution over $\mathbb{R}^3$. Here are two cases:

Case 1: $D(P^*) = \underset{x\sim D}{\mathrm{P}}[x \in P^*] \le \varepsilon$, in this case:

$$L_{h^*,D}(h_S) = \underset{x\sim D}{\mathrm{P}}[h_S(x) \ne h^*(x)] = \underset{x\sim D}{\mathrm{P}}[x \in P^* \backslash P_S] \le \underset{x\sim D}{\mathrm{P}}[x \in P^*] \le \varepsilon \Rightarrow \underset{S\sim D^m}{\mathrm{P}}\left[L_{h^*,D}(h_S) \le \varepsilon\right] = 1 > \delta \in (0,1)\ (1)$$

Case 2: $D(P^*) = \underset{x\sim D}{\mathrm{P}}[x \in P^*] > \varepsilon$.

Let $\begin{cases} R_1 = [a_1^*, a_1^* + c] \times [a_2^*, a_2^* + r^*] \\ R_2 = [a_1^* + r^* - c, a_1^* + r^*] \times [a_2^*, a_2^* + r^*] \\ R_3 = [a_1^*, a_1^* + r^*] \times [a_2^*, a_2^* + c] \\ R_4 = [a_1^*, a_1^* + r^*] \times [a_2^*, a_2^* + r^* - c] \end{cases}$ $such\ that\ D(\mathrm{R_i})_{i \in \overline{1,4}} = \frac{\varepsilon}{4}, where\ c \in \mathbb{R}_+^* \Leftrightarrow \underset{x\sim D}{\mathrm{P}}[x \in R_i]_{i\in\overline{1,4}} = \frac{\varepsilon}{4}$

Case 2.1: $P_S \cap R_i \ne \emptyset\ \forall i \in \overline{1,4}$

$$L_{h^*,D}(h_S) = \underset{x\sim D}{\mathrm{P}}\left[h_{S(x)} \ne h^*(x)\right] = \underset{x\sim D}{\mathrm{P}}[x \in P^* \backslash P_S] \le \underset{x\sim D}{\mathrm{P}}\left[x \in \bigcup_{i=1}^{4} R_i\right] \le \sum_{i=1}^{4} \underset{x\sim D}{\mathrm{P}}[x \in R_i] = 4\frac{\varepsilon}{4} = \varepsilon \Rightarrow \underset{S\sim D^m}{\mathrm{P}}\left[L_{h^*,D}(h_S) \le \varepsilon\right] = 1 > \delta \in (0,1)\ (2)$$

Case 2.1: $\exists i \in \overline{1,4}\ s.t. P_S \cap R_i = \emptyset$ then we have to enlarge S.

Let $F_i = \{S\sim D^m |\ P_S \cap R_i = \emptyset\}$

$$\underset{S\sim D^m}{\mathrm{P}}\left[L_{h^*,D}(h_S) > \varepsilon\right] = \underset{S\sim D^m}{\mathrm{P}}\left[\bigcup_{i=1}^{4} F_i\right] \le \sum_{i=1}^{4} \underset{x\sim D}{\mathrm{P}}[x \in F_i]$$

$$= the\ probability\ that\ P_S\ will\ not\ intersect\ \underbrace{R_i}_{\varepsilon}\ (S\ is\ i.i.d.)_{i\in\overline{1,4}}$$

$$= 4\left(1 - \frac{\varepsilon}{4}\right)^m (1 - x \le e^{-x}) \le 4e^{-\frac{\varepsilon}{4}m} < \delta \Rightarrow m > \frac{4ln\frac{4}{\delta}}{\varepsilon}$$

$$\boxed{m \geq m_{\mathcal{H}}(\varepsilon,\delta) = \left\lceil \frac{4ln\frac{4}{\delta}}{\varepsilon} \right\rceil} (3)$$

$if\ W > H$:

We can see 4 areas of interest:

There are 2 areas that cause error: P^*/P_S (false negative) because here are the only points that have the label 1 but are labeled with 0 by h_S, and P_S/P^*(false positive) because here are the only points that have the label 0 but are labeled with 1 by h_S. In addition the points in $\mathbb{R}^3/(P_S \cup P^*)$ have label 0 and are labeled with 0 by h_S ✓, and the points in $P_S \cap P^*$ have label 1 and are labeled with 1 by h_S ✓.

Let's fixe $\varepsilon \in (0,1)$ and $\delta \in (0,1)$ and let D distribution over $\mathbb{R}^3$. Here are two cases:

Case 1: $D(P^*) = \underset{x\sim D}{\mathrm{P}}[x \in P^*] \leq \frac{3\varepsilon}{4}$ and $\underset{x\sim D}{\mathrm{P}}[x \in P_S \backslash P^*] \leq \frac{\varepsilon}{4}$, in this case:

$$L_{h^*,D}(h_S) = \underset{x\sim D}{\mathrm{P}}[h_S(x) \neq h^*(x)] = \underset{x\sim D}{\mathrm{P}}[x \in P^* \backslash P_S] + \underset{x\sim D}{\mathrm{P}}[x \in P_S \backslash P^*] \leq \varepsilon \Rightarrow \underset{S\sim D^m}{\mathrm{P}}[L_{h^*,D}(h_S) \leq \varepsilon] = 1 > \delta \in (0,1) \ (4)$$

Case 2: $D(P^*) = \underset{x\sim D}{\mathrm{P}}[x \in P^*] > \frac{3\varepsilon}{4}$ or $\underset{x\sim D}{\mathrm{P}}[x \in P_S \backslash P^*] > \frac{\varepsilon}{4}$.

Let $\begin{cases} R_1 = [a_1^*, a_1^* + c] \times [a_2^*, a_2^* + r^*] \\ R_2 = [a_1^* + r^* - c, a_1^* + r^*] \times [a_2^*, a_2^* + r^*] \\ R_3 = [a_1^*, a_1^* + r^*] \times [a_2^*, a_2^* + c] \\ R_4 = [a_1, a_1 + r] \times [a_2^* + r^*, a_2 + r + d] \end{cases}$ $such\ that\ D(\mathrm{R_i})_{i\in\overline{1,4}} = \frac{\varepsilon}{4}, where\ c, d \in \mathbb{R}_+^* \Leftrightarrow \underset{x\sim D}{\mathrm{P}}[x \in R_i]_{i\in\overline{1,4}} = \frac{\varepsilon}{4}$

Case 2.1: $P_S \cap R_i \neq \emptyset\ \forall i \in \overline{1,4}$

$$L_{h^*,D}(h_S) = \underset{x\sim D}{\mathrm{P}}\left[h_{S(x)} \neq h^*(x)\right] = \underset{x\sim D}{\mathrm{P}}[x \in P^* \backslash P_S] + \underset{x\sim D}{\mathrm{P}}[x \in P_S \backslash P^*] \leq \underset{x\sim D}{\mathrm{P}}\left[x \in \bigcup_{i=1}^{4} R_i\right]$$

$$\leq \sum_{i=1}^{4} \underset{x\sim D}{\mathrm{P}}[x \in R_i] = 4\frac{\varepsilon}{4} = \varepsilon \Rightarrow \underset{S\sim D^m}{\mathrm{P}}[L_{h^*,D}(h_S) \leq \varepsilon] = 1 > \delta \in (0,1) \ (5)$$

Case 2.1: $\exists i \in \overline{1,4\ s.t.} P_S \cap R_i = \emptyset$ then we have to enlarge S.

Let $F_i = \{S\sim D^m |\ P_S \cap R_i = \emptyset\}$

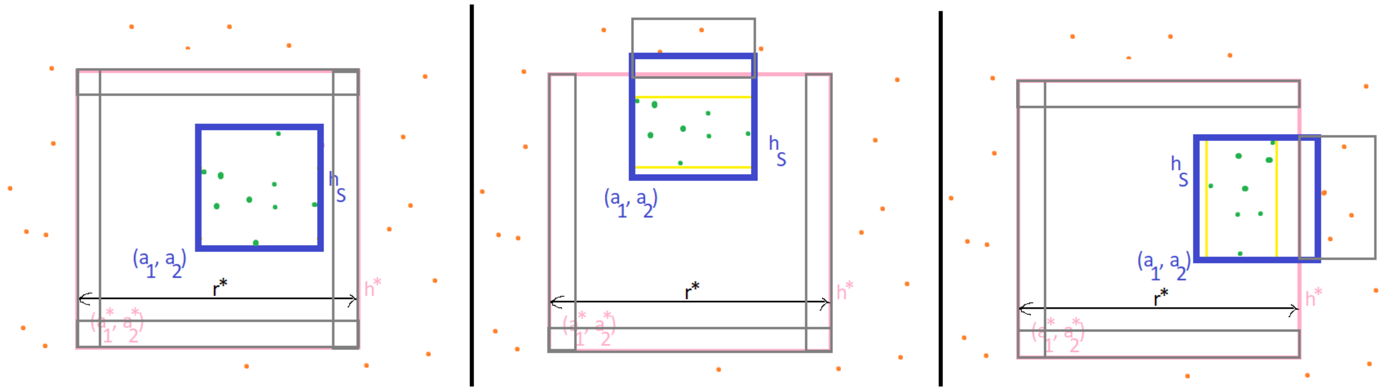
$$\underset{S\sim D^m}{\mathrm{P}}[L_{h^*,D}(h_S) > \varepsilon] = \underset{S\sim D^m}{\mathrm{P}}\left[\bigcup_{i=1}^{4} F_i\right] \leq \sum_{i=1}^{4} \underset{x\sim D}{\mathrm{P}}[x \in F_i]$$

$$= the\ probability\ that\ P_S\ will\ not\ intersect\ \underbrace{R_i}_{\varepsilon}\ (S\ is\ i.i.d.)_{i\in\overline{1,4}}$$

$$= 4\left(1 - \frac{\varepsilon}{4}\right)^m (1 - x \leq e^{-x}) \leq 4e^{-\frac{\varepsilon}{4}m} < \delta \Rightarrow m > \frac{4ln\frac{4}{\delta}}{\varepsilon}$$

$$\boxed{m \geq m_{\mathcal{H}}(\varepsilon,\delta) = \left\lceil \frac{4ln\frac{4}{\delta}}{\varepsilon} \right\rceil} (6)$$

$$Case\ W < H\ is\ analogus\ to\ case\ W > H (7)$$



It can be seen that in each case we have 4 error zones

$$(1),(2),(3),(4),(5),(6),(7) \Rightarrow \mathcal{H} \text{ can be } (\varepsilon,\delta) - \text{PAC learned by A (defined above)}$$

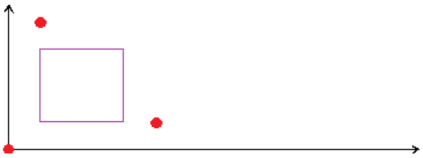
$$\textit{and has } \text{sample complexity } \boxed{m_{\mathcal{H}}(\varepsilon,\delta) = \left\lceil \frac{4ln\frac{4}{\delta}}{\varepsilon} \right\rceil} \blacksquare$$

Solution problem 4 c:

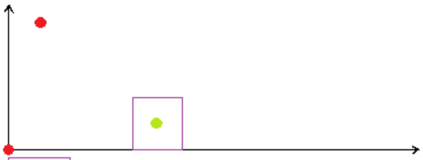
We will prove that $\begin{matrix} VCdim(\mathcal{H}) \geq 3 & (1) \\ VCdim(\mathcal{H}) < 4 & (2) \end{matrix}$ which will lead to $VCdim(\mathcal{H}) = 3$.

In the first phase we will prove (1), that is we must find a set A such that $\mathcal{H}$ shatters A. Let $A = \{(0,0),(1,5),(5,1)\}$ a set with three points on the real axis. Next, for any labeling on the points in A, we must find a classifier $h_A \in \mathcal{H}$ such that the error on set A of the classifier is 0, equivalent $L_{A(h_A)} = 0$.

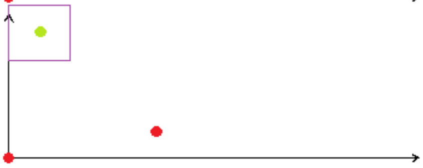
For the label $\begin{cases} (0,0) = 0 \\ (1,5) = 0 \\ (5,1) = 0 \end{cases}$. $h_{(a_1,a_2,r)}$ with $\begin{cases} a_1 = 1 \\ a_2 = 1 \\ r = 3 \end{cases}$ has 0 error.



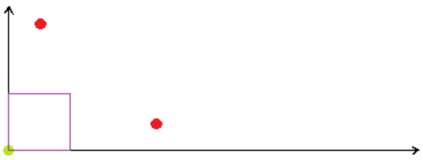
For the label $\begin{cases} (0,0) = 0 \\ (1,5) = 0 \\ (5,1) = 1 \end{cases}$. $h_{(a_1,a_2,r)}$ with $\begin{cases} a_1 = 4 \\ a_2 = 0 \\ r = 2 \end{cases}$ has 0 error.



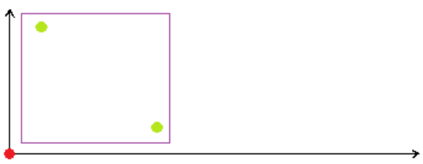
For the label $\begin{cases} (0,0) = 0 \\ (1,5) = 1 \\ (5,1) = 0 \end{cases}$. $h_{(a_1,a_2,r)}$ with $\begin{cases} a_1 = 0 \\ a_2 = 4 \\ r = 2 \end{cases}$ has 0 error.



For the label $\begin{cases} (0,0) = 1 \\ (1,5) = 0 \\ (5,1) = 0 \end{cases}$. $h_{(a_1,a_2,r)}$ with $\begin{cases} a_1 = 0 \\ a_2 = 0 \\ r = 2 \end{cases}$ has 0 error.

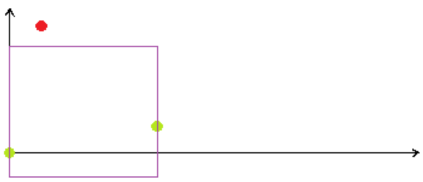


For the label $\begin{cases} (0,0) = 0 \\ (1,5) = 1 \\ (5,1) = 1 \end{cases}$. $h_{(a_1,a_2,r)}$ with $\begin{cases} a_1 = 0,5 \\ a_2 = 0,5 \\ r = 5 \end{cases}$ has 0 error.

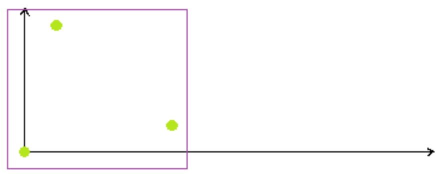


For the label $\begin{cases} (0,0) = 1 \\ (1,5) = 1 \\ (5,1) = 0 \end{cases}$. $h_{(a_1,a_2,r)}$ with $\begin{cases} a_1 = -1 \\ a_2 = 0 \\ r = 5 \end{cases}$ has 0 error.

For the label $\begin{cases} (0,0) = 1 \\ (1,5) = 0 \\ (5,1) = 1 \end{cases}$. $h_{(a_1,a_2,r)}$ with $\begin{cases} a_1 = -1 \\ a_2 = 0 \\ r = 5 \end{cases}$ has 0 error.



For the label $\begin{cases}(0,0) = 1 \\ (1,5) = 1 \\ (5,1) = 1\end{cases}$. $h_{(a_1,a_2,r)}$ with $\begin{cases}a_1 = -0.5 \\ a_2 = -0.5 \\ r = 6\end{cases}$ has 0 error.



So for any labeling of the points in A, that is $2^{|A|} = 2^3 = 8$ possible labels (tested above), we found one classifier to learn that labeling. So (1) is proven $\Rightarrow VCdim(\mathcal{H}) \geq |A| = 3$ (3)

To prove (2) it is necessary that $\forall A, |A| = 4$, $\mathcal{H}$ does not shatter A. Let $A = \{a_1, a_2, a_3, a_4 |\ a_i = (x_i, y_i) \in \mathbb{R}^2, i \in \overline{1,4}\}$. Because A is a set $\Rightarrow a_i \neq a_j, \forall i \neq j \in \overline{1,4}$. There are two cases below: the points are degenerate (there are collinear points $\Leftrightarrow \exists a_l, a_m, a_n \in A\ sunch\ that\ a_l - a_m - a_n$) and there are no collinear points ($\Leftrightarrow \nexists a_l, a_m, a_n \in A\ sunch\ that\ a_l - a_m - a_n$).

Case 1 - points are degenerate:

there are collinear points $\Leftrightarrow \exists a_l, a_m, a_n \in A\ sunch\ that\ a_l - a_m - a_n$. Without restricting the generality we can consider $l < m < n$. Next we will show that the label $\begin{cases}a_l = 1 \\ a_m = 0 \\ a_n = 1 \\ a_q = random\end{cases}$, $q \in \overline{1,4}/\{l,m,n\}$ cannot be learned. For this we will use Reductio ad Absurdum Method: Suppose that: $\exists h_P \in \mathcal{H}$ such that $L_{A(h_A)} = 0$

$$\Rightarrow \exists a_{1_P}, a_{2_P}, r_P\ such\ that \begin{cases}a_l = 1 \\ a_m = 0 \\ a_n = 1\end{cases} \Rightarrow \begin{cases} a_l \in [a_{1_P}, a_{1_P} + r_P] \times [a_{2_P}, a_{2_P} + r_P] \Rightarrow \begin{cases} x_l \in [a_{1_P}, a_{1_P} + r_P] \\ y_l \in [a_{2_P}, a_{2_P} + r_P]\end{cases} & (4) \\ a_m \notin [a_{1_P}, a_{1_P} + r_P] \times [a_{2_P}, a_{2_P} + r_P] & (5) \\ a_n \in [a_{1_P}, a_{1_P} + r_P] \times [a_{2_P}, a_{2_P} + r_P] \Rightarrow \begin{cases} x_l \in [a_{1_P}, a_{1_P} + r_P] \\ y_l \in [a_{2_P}, a_{2_P} + r_P]\end{cases} & (6) \end{cases}$$

$$a_l - a_m - a_n \Rightarrow \begin{cases} x_l < x_m < x_n \\ y_l < y_m < y_n \end{cases} \Rightarrow a_m \in [x_l, x_n] \times [y_l, y_n] \Bigg|_{\substack{(4) \\ (6)}} \Rightarrow a_m \in [a_{1_P}, a_{1_P} + r_P] \times [a_{2_P}, a_{2_P} + r_P] \Bigg|_{(5)}$$

$$\Rightarrow \perp \Rightarrow \nexists h_P \in \mathcal{H} \text{ such that } L_{A(h_A)} = 0 \quad (7)$$

Case 2 - there are no collinear points:

$\nexists a_l, a_m, a_n \in A\ sunch\ that\ a_l - a_m - a_n \Rightarrow \Delta a_l\, a_m a_n\ is\ undegenerated$. If $a_q \in int(\Delta a_l\, a_m a_n), q \in \overline{1,4}/\{l,m,n\}$, then the label $\begin{cases}a_l = 1 \\ a_m = 1 \\ a_n = 1 \\ a_q = 0\end{cases}$ it cannot be learned. For this we will use Reductio ad Absurdum Method: Suppose that: $\exists h_P \in \mathcal{H}$ such that $L_{A(h_A)} = 0$

$$\Rightarrow \exists a_{1_P}, a_{2_P}, r_P \text{ such that } \begin{cases} a_l = 1 \\ a_m = 1 \\ a_n = 1 \\ a_q = 0 \end{cases} \Rightarrow \begin{cases} a_l \in [a_{1_P}, a_{1_P} + r_P] \times [a_{2_P}, a_{2_P} + r_P] & (8) \\ a_m \in [a_{1_P}, a_{1_P} + r_P] \times [a_{2_P}, a_{2_P} + r_P] & (9) \\ a_n \in [a_{1_P}, a_{1_P} + r_P] \times [a_{2_P}, a_{2_P} + r_P] & (10) \\ a_q \notin [a_{1_P}, a_{1_P} + r_P] \times [a_{2_P}, a_{2_P} + r_P] & (11) \end{cases}$$

$$\left. \begin{array}{c} (8),(9),(10) \Rightarrow \Delta a_l\, a_m a_n \in [a_{1_P}, a_{1_P} + r_P] \times [a_{2_P}, a_{2_P} + r_P] \\ (11) \\ a_q \in int(\Delta a_l\, a_m a_n) \end{array} \right| \Rightarrow \perp \Rightarrow a_q \notin int(\Delta a_l\, a_m a_n)$$

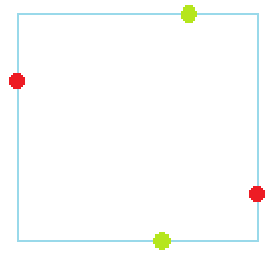
Furthermore, the points from A describe a convex quadrilateral. Otherwise one of the points is inside the polygon given by the rest of the dot (i.e. triangle), and a demonstration that follows analogous to the previous one leads us to the previous statement (the points from A describe a convex quadrilateral).

We now have well-defined notions of right, left, bottom and top. And let: $\begin{cases} a_1 = \textit{the rightmost point} \\ a_2 = \textit{the highest point} \\ a_3 = \textit{the leftmost point} \\ a_4 = \textit{the lowest point} \end{cases}$

$$\Leftrightarrow \begin{cases} \sphericalangle(a_1) < \sphericalangle(a_2) < \sphericalangle(a_3) < \sphericalangle(a_4) \\ a_1 a_2 a_3 a_4 \text{ convex quadrilateral} \end{cases},$$

where $\sphericalangle(.)$ the polar angle of a point in relation to G - center of gravity of the polygon $a_1 a_2 a_3 a_4$.

Let $\begin{cases} W \stackrel{\text{def}}{=} |x_1 - x_3| \\ H \stackrel{\text{def}}{=} |y_2 - y_4| \end{cases}$. Here are 3 cases:



If $H = W$:

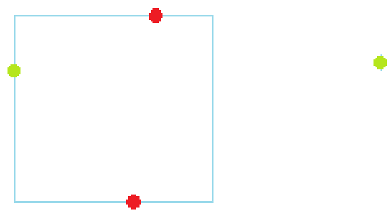
Next we will show that the label (0,1,0,1) cannot be learned. For this we will use Reductio ad Absurdum Method: Suppose that: $\exists h_P \in \mathcal{H}$ such that $L_{A(h_A)} = 0$

$$\Rightarrow \exists a_{1_P}, a_{2_P}, r_P \text{ such that } \begin{cases} a_1 = 0 \\ a_2 = 1 \\ a_3 = 0 \\ a_4 = 1 \end{cases} \Rightarrow \begin{cases} a_1 \notin [a_{1_P}, a_{1_P} + r_P] \times [a_{2_P}, a_{2_P} + r_P] \Rightarrow x_1 > a_{1_P} + r_P & (8) \\ a_2 \in [a_{1_P}, a_{1_P} + r_P] \times [a_{2_P}, a_{2_P} + r_P] & (9) \\ a_3 \notin [a_{1_P}, a_{1_P} + r_P] \times [a_{2_P}, a_{2_P} + r_P] \Rightarrow x_3 < a_{1_P} & (10) \\ a_4 \in [a_{1_P}, a_{1_P} + r_P] \times [a_{2_P}, a_{2_P} + r_P] & (11) \end{cases}$$

$$\left. \begin{array}{c} (9),(10) \Rightarrow H \stackrel{\text{def}}{=} |y_2 - y_4| \in [0, r_P] \\ (8),(9) \Rightarrow W \stackrel{\text{def}}{=} |x_1 - x_3| > r_P \\ H = W \end{array} \right| \Rightarrow \perp \Rightarrow \nexists h_P \in \mathcal{H} \text{ such that } L_{A(h_A)} = 0 \quad (12)$$

If $H < W$:

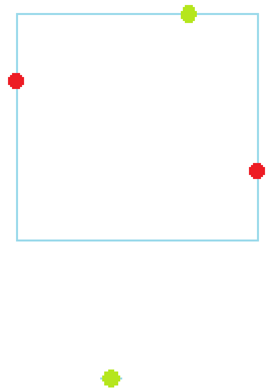
Next we will show that the label (1,0,1,0) cannot be learned. For this we will use Reductio ad Absurdum Method: Suppose that: $\exists h_P \in \mathcal{H}$ such that $L_{A(h_A)} = 0$



$$\Rightarrow \exists a_{1_P}, a_{2_P}, r_P \text{ such that } \begin{cases} a_1 = 1 \\ a_2 = 0 \\ a_3 = 1 \\ a_4 = 0 \end{cases} \Rightarrow \begin{cases} a_1 \in [a_{1_P}, a_{1_P} + r_P] \times [a_{2_P}, a_{2_P} + r_P] & (13) \\ a_2 \notin [a_{1_P}, a_{1_P} + r_P] \times [a_{2_P}, a_{2_P} + r_P] \Rightarrow y_2 > a_{2_P} + r_P & (14) \\ a_3 \in [a_{1_P}, a_{1_P} + r_P] \times [a_{2_P}, a_{2_P} + r_P] & (15) \\ a_4 \notin [a_{1_P}, a_{1_P} + r_P] \times [a_{2_P}, a_{2_P} + r_P] \Rightarrow y_4 < a_{2_P} & (16) \end{cases}$$

$$\left.\begin{array}{c}(14),(16) \Rightarrow H \stackrel{\text{def}}{=} |y_2 - y_4| > r_P \\ (13),(15) \Rightarrow W \stackrel{\text{def}}{=} |x_1 - x_3| \in [0, r_P] \\ H < W\end{array}\right| \Rightarrow \perp \Rightarrow \nexists h_P \in \mathcal{H} \text{ such that } L_{A(h_A)} = 0 \quad (17)$$

If $H > W$:



Next we will show that the label (0,1,0,1) cannot be learned. For this we will use Reductio ad Absurdum Method: Suppose that: $\exists h_P \in \mathcal{H}$ such that $L_{A(h_A)} = 0$

$$\Rightarrow \exists a_{1_P}, a_{2_P}, r_P \; such\ that \begin{cases} a_1 = 0 \\ a_2 = 1 \\ a_3 = 0 \\ a_4 = 1 \end{cases}$$

$$\Rightarrow \begin{cases} a_1 \notin \left[a_{1_P}, a_{1_P} + r_P\right] \times \left[a_{2_P}, a_{2_P} + r_P\right] \Rightarrow x_1 > a_{1_P} + r_P & (18) \\ a_2 \in \left[a_{1_P}, a_{1_P} + r_P\right] \times \left[a_{2_P}, a_{2_P} + r_P\right] & (19) \\ a_3 \notin \left[a_{1_P}, a_{1_P} + r_P\right] \times \left[a_{2_P}, a_{2_P} + r_P\right] \Rightarrow x_3 < a_{1_P} & (20) \\ a_4 \in \left[a_{1_P}, a_{1_P} + r_P\right] \times \left[a_{2_P}, a_{2_P} + r_P\right] & (21) \end{cases}$$

$$\left.\begin{array}{c}(19),(21) \Rightarrow H \stackrel{\text{def}}{=} |y_2 - y_4| \in [0, r_P] \\ (18),(20) \Rightarrow W \stackrel{\text{def}}{=} |x_1 - x_3| > r_P \\ H > W\end{array}\right| \Rightarrow \perp \Rightarrow \nexists h_P \in \mathcal{H} \text{ such that } L_{A(h_A)} = 0 \quad (22)$$

$$(7),(12),(17),(22) \Rightarrow \nexists h_P \in \mathcal{H} \text{ such that } L_{A(h_A)} = 0 \Rightarrow VCdim(\mathcal{H}) < |A| = 4 \; (23)$$

$$(3),(23) \Rightarrow \boxed{VCdim(\mathcal{H}) = 3} \blacksquare$$

Bonus Problem. Compute the VC-dimension of the class of convex d-gons (convex polygons with exactly d sides) in the plane. Provide a detailed proof of your result.

Solution Bonus Problem:

$$\mathcal{H} = \left\{h_{(a_1,a_2,\dots,a_d)}: \mathbb{R}^2 \to \{0,1\} \mid h_{(a_1,a_2,\dots,a_d)}(x) = \begin{cases}1, if\ x \in CH(a_1,a_2,\dots,a_d)\\ 0, otherwise\end{cases} \quad a_i \in \mathbb{R}^2, \right.$$
$$\left. \text{i} \in \overline{1,d}, CH - convex\ hull, \nexists j \in \overline{1,d}\ s.t.\ CH(a_1,\dots a_{j-1},a_{j+1},\dots a_d) = CH(a_1,a_2,\dots,a_d)\right\}$$

Before starting the demonstration, we will make a series of statements:

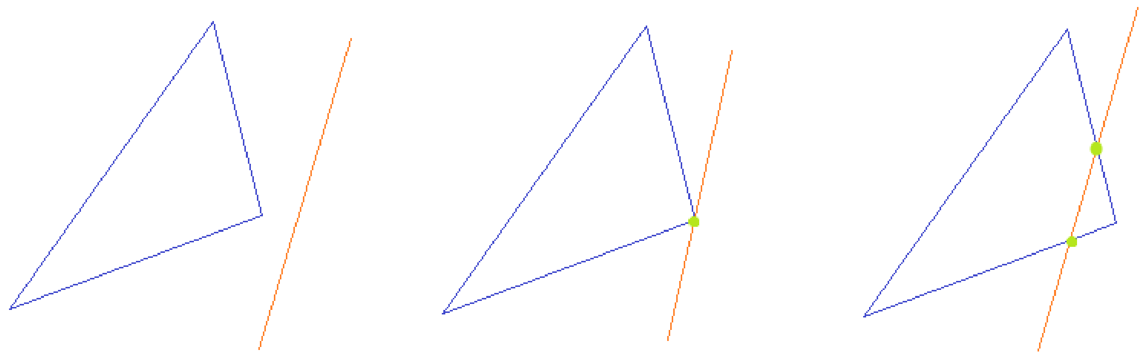
1. $|d \cap LP(A)| \le 2, \forall d\ line\ in\ plane, \forall A\ convex\ polygon$, where $LP(A) \stackrel{\text{def}}{=} polygonal\ line\ of\ A = (a_1,a_2,\dots,a_n) \stackrel{\text{def}}{=} \cup_{i=1}^{n-1}[a_i,a_{i+1}] \cup [a_n,a_1], n > 2, d \notin \left\{a_ia_{i+1}{}_{i\in\overline{1,d-1}}, a_da_1\right\}$ $\quad (A1)$
2. $L_{\{A,B\}}(h) = 0 \Leftrightarrow |[A,B] \cap m| = 1, \forall A,B \in \mathbb{R}^2, label(A) \neq label(B), where\ h_{(a_1,a_2,\dots,a_d)}(x) = \begin{cases}1, if\ x \in CH(a_1,a_2,\dots,a_d)\\ 0, otherwise\end{cases}, CH - convex\ hull, \nexists j \in \overline{1,d}\ s.t.\ CH(a_1,\dots a_{j-1},a_{j+1},\dots a_d) = CH(a_1,a_2,\dots,a_d), m \in \left\{[a_i,a_{i+1}]_{i\in\overline{1,d-1}}, [a_d,a_1]\right\}\ (edge\ of\ the\ classifier\ \ d-gon)$ $\quad (A2)$

Demonstration A1:

We will use the Principle of Mathematical Induction:

Let $P[n] \stackrel{\text{def}}{=}$ "$|d \cap LP(A)| \le 2, \forall d\ line\ in\ plane, \forall A = (a_1,a_2,\dots,a_n)\ convex\ n-gon,\ d \notin \left\{a_ia_{i+1}{}_{i\in\overline{1,n-1}}, a_na_1\right\}$"

$P[3] =$ "$|d \cap LP(A)| \le 2, \forall d\ line\ in\ plane, \forall A = (a_1,a_2,a_3)\ triangle, d \notin \{a_1a_2, a_2a_3, a_3a_1\}$" is trivially true, because the only possibilities of the existence of a line with respect to a triangle are:



We assume that $P[k]$="$|d \cap LP(A)| \le 2, \forall d\ line\ in\ plane, \forall A = (a_1,a_2,\dots,a_k)\ convex\ k-gon,\ d \notin \left\{a_ia_{i+1}{}_{i\in\overline{1,k-1}}, a_ka_1\right\}$" is true.

Next we prove that $P[k] \to P[k+1]$.

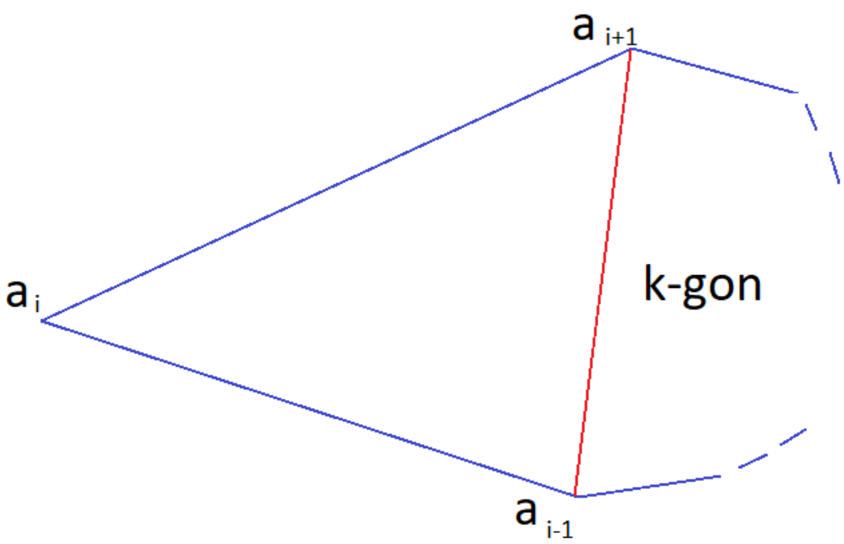
$P[k+1]$="$|d \cap LP(A)| \le 2, \forall d\ line\ in\ plane, \forall A = (a_1,a_2,\dots,a_{k+1})\ convex\ k+1-gon,\ d \notin \left\{a_ia_{i+1}{}_{i\in\overline{1,k}}, a_{k+1}a_1\right\}$".

Let $A = (a_1,a_2,\dots,a_{k+1})\ convex\ k+1-gon \Rightarrow \forall (a_1,\dots,a_{i-1},a_{i+1},\dots,a_{k+1}) \subseteq int(\measuredangle i), i \in \overline{1,k+1}$ $\quad (1)$

Let's fix $i \in \overline{1,k+1}$. Convention: $a_{k+2} = a_1$

Let $\begin{cases} Z_1 \stackrel{\text{def}}{=} CH(a_{i-1},a_i,a_{i+1}) \\ Z_2 \stackrel{\text{def}}{=} CH(a_1,\dots,a_{i-1},a_{i+1},\dots,a_{k+1})\end{cases}$ and let $d\ line\ in\ plane, d\ \notin \left\{a_ia_{i+1}{}_{i\in\overline{1,k}}, a_{k+1}a_1\right\}$

There are 4 cases:



Case 1: $d \cap Z_1 = \emptyset \ and \ d \cap Z_2 \neq \emptyset$

$$\left.\begin{array}{c}(a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_{k+1}) \ is \ k - gon \\ P[k] \text{ is true.}\end{array}\right| \Rightarrow \left|d \cap LP\big((a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_{k+1})\big)\right| \leq 2 \ \ (2)$$

Case 2: $d \cap Z_2 = \emptyset \ and \ d \cap Z_1 \neq \emptyset$

Case 2 is reduced to $P[3]$, which is demonstrated above. (3)

Case 3: $d \cap a_{i-1}a_{i+1} \neq \emptyset$

$$\left.\begin{array}{c} \mathrm{A} = (a_1, a_2, \dots, a_{k+1}) \ convex \ k+1-gon \Rightarrow \left.\begin{array}{c}{\scriptstyle (a_{i-1}, a_i, a_{i+1}) \ triangle} \\ {\scriptstyle d \cap a_{i-1}a_{i+1} \neq \emptyset}\end{array}\right| \Rightarrow \left.\begin{cases} d \cap a_i a_{i+1} \neq \emptyset \\ \qquad or \\ d \cap a_{i-1} a_i \neq \emptyset \end{cases}\right| \\ (2) \Rightarrow \left|d \cap \big(LP\big((a_1, \dots, a_{i-1}, a_{i+1}, \dots, a_{k+1})\big)/[a_{i-1}a_{i+1}]\big)\right| \leq 1 \end{array}\right| \Rightarrow$$

$$\Rightarrow \left|d \cap \big(LP(a_1, \dots, a_{i-1}, a_i, a_{i+1}, \dots, a_{k+1})\big)\right| \leq 1 + 1 = 2 \ \ (4)$$

Case 4: $d \cap (Z_1 \cup Z_2) = \emptyset$

$$d \cap (Z_1 \cup Z_2) = \emptyset \Leftrightarrow d \cap A = \emptyset \Leftrightarrow |d \cap LP(A)| = 0 \ (5)$$

$(2), (3), (4), (5) \Rightarrow |d \cap LP(A)| \leq 2, \forall d \ line \ in \ plane, \forall A = (a_1, a_2, \dots, a_{k+1}) \ convex \ k+1-gon,$
$d \notin \left\{a_i a_{i+1}{}_{i \in \overline{1,k}}, a_{k+1}a_1\right\} \Rightarrow P[k+1] \ is \ true$

So $P[k] \rightarrow P[k+1] \Rightarrow P[n] \ is \ true \ \forall n > 2, n \in \mathbb{N}$. So the statement (A1) is proven.

Demonstration A2:

Let $A, B \in \mathbb{R}^2, label(A) \neq label(B)$.

If $A, B \in CH(a_1, a_2, \dots, a_d)$:

$$L_{\{A,B\}}(h) = 0 \ \Leftrightarrow \begin{cases} label(A) = 1 \\ label(B) = 1 \end{cases}, but \ label(A) \neq label(B) \ \Rightarrow \perp \Rightarrow A, B \notin CH(a_1, a_2, \dots, a_d) \ \ (1)$$

If $A, B \notin CH(a_1, a_2, \dots, a_d)$:

$$L_{\{A,B\}}(h) = 0 \ \Leftrightarrow \begin{cases} label(A) = 0 \\ label(B) = 0 \end{cases}, but \ label(A) \neq label(B) \ \Rightarrow \perp \Rightarrow A, B \in CH(a_1, a_2, \dots, a_d) \ \ (2)$$

$$(1),(2) \Rightarrow \begin{cases} A \in CH(a_1, a_2, \dots, a_d) \text{ and } B \notin CH(a_1, a_2, \dots, a_d) \\ \text{or} \\ B \in CH(a_1, a_2, \dots, a_d) \text{ and } A \notin CH(a_1, a_2, \dots, a_d) \end{cases} \Bigg| \Rightarrow$$

$$\Rightarrow \begin{cases} L_{\{A,B\}}(h) = 0 \\ \exists m \in \{[a_i, a_{i+1}]_{i\in\overline{1,d-1}}, [a_d, a_1]\}\ s.t. |[A,B] \cap m| = 1 \end{cases}$$

So the statement (A2) is proven.

We will prove that $\begin{matrix} VCdim(\mathcal{H}) \geq 2d+1 & (3) \\ VCdim(\mathcal{H}) < 2d+2 & (4) \end{matrix}$ which will lead to $VCdim(\mathcal{H}) = 2d+1$.

In the first phase we will prove (3), that is we must find a set A such that $\mathcal{H}$ shatters A. Let $A = \{A_i | A_i \in \mathbb{R}^2, \mathrm{i} \in \overline{1, 2d+1}, \textit{the vertices of a regular polygon}\}$ a set with 2d+1 points on the real axis. Next, for any labeling on the points in A, we must find a classifier $h_A \in \mathcal{H}$ such that the error on set A of the classifier is 0, equivalent $L_{A(h_A)} = 0$.

Convention: $A_{2d+2} = A_1$

We will first define the notion of cluster:

$$Cluster(A_i) = \begin{cases} \{A_i\}, & \textit{if } label(A_{i-1}) \neq label(A_i) \neq label(A_{i+1}) \\ \{A_i\} \cup Cluster(A_{i-1}), & \textit{if } label(A_{i-1}) = label(A_i) \neq label(A_{i+1}) \\ \{A_i\} \cup Cluster(A_{i+1}), & \textit{if } label(A_{i+1}) = label(A_i) \neq label(A_{i-1}) \\ \{A_i\} \cup Cluster(A_{i-1}) \cup Cluster(A_{i+1}), & \textit{if } label(A_{i-1}) = label(A_i) = label(A_{i+1}) \end{cases}$$

Note: There are no neighboring clusters labeled 0, as they would have merged when calculating the clusters sets. And no neighboring clusters with label 1 for the same reason. The only possibilities are: cluster 1 next to cluster 0.

Let $(l_i)_{i\in\overline{1,2d+1}}$ be a labeling of the points in A.

Let the following algorithm:

1. List $= \emptyset$
2. Determine the clusters until they stabilize.
3. Sort the clusters by the polar angle in relation to the center of gravity of A (possible from the definition of clusters - neighboring nodes)
4. for $i \in \overline{1, \#clusters - 1}$:

    if $label(Cluster_i = 1)$:

    $\mathrm{A}^i_{max} = \textit{the highest index from } Cluster_i$

    $\mathrm{A}^{i+1}_{min} = \textit{the lowest index from } Cluster_{i+1}$

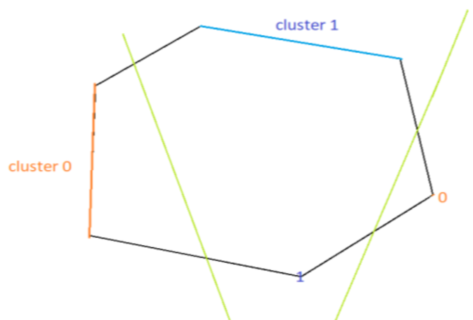
    $\mathrm{A}^{i+1}_{max} = \textit{the highest index from } Cluster_{i+1}$

    $\mathrm{A}^{i+2}_{min} = \textit{the lowest index from } Cluster_{i+2}$

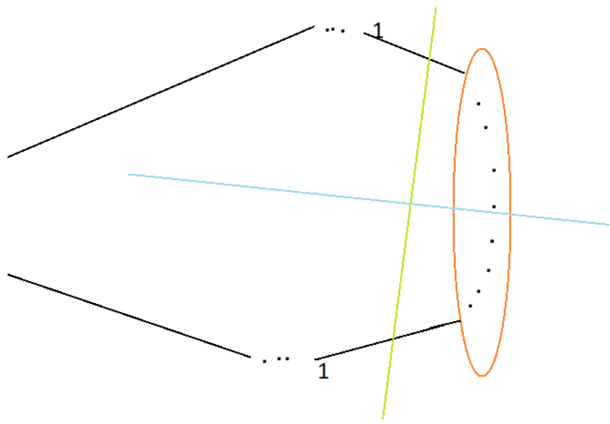
    $\textit{Let } d \textit{ line in plane s.t. } d \cap [\mathrm{A}^i_{max}, \mathrm{A}^{i+1}_{min}] \neq \emptyset \textit{ and } d \cap [\mathrm{A}^{i+1}_{max}, \mathrm{A}^{i+2}_{min}] \neq \emptyset,$

    $d \nparallel e, e \in List$

    List $\leftarrow d$
5. returns the list of intersection points of the lines in the List



Of course there is a possibility that the lines do not intersect and the points at infinity will be considered as points of intersection. Correctness does not change. An example on the right, the classifier is inside the angle determined by the green lines.



Correctness:

According to the construction of the algorithm we have $L_{A(h_A)} = 0$ (5).

It must also be shown that the result is d-gon

The right scenario cannot happen, i.e. there are right intersections inside A. Proof: we assume that such an intersection exists. Let d be a line given by the algorithm $\Rightarrow$ on one side of the line d there is a single cluster $Cluster_d$ and on the other side the rest of the polygon A $(Cluster_{d-1}\ and\ Cluster_{d+1} with\ different\ label\ than\ Cluster_d)$. Let $e, e \cap d \cap A \neq \emptyset$ and line e is generated by the algorithm $\Rightarrow e \cap LP(Cluster_d)$
$\Rightarrow \exists\ a\ label\ change\ in\ Cluster_d$. But $Cluster_d$ is cluster $\Rightarrow \perp$
$\Rightarrow \nexists$ intersections inside A $\Rightarrow \exists\ intersections\ outside\ A$ (6).

The resulting polygon is convex in construction but also due to the convexity of A.

Furthermore, $|A| = 2d + 1 \Rightarrow \#clusters \leq 2d + 1$. The maximum number of clusters is reached when the number of clusters alternates. $\#clusters = 2d$ (due to the odd number of nodes, two borders with equal labels will be next to each other, i.e. in the same cluster.) $\Rightarrow \#clusters \leq 2d$ (7)

$$(7), (A1), (A2) \Rightarrow the\ maximum\ number\ of\ nodes\ of\ the\ classifier\ polygon\ is\ d\ (8)$$

If the number of nodes is less than d then the polygon can artificially increase at d-gon.

(5), (6), (7), (8) $\Rightarrow$ For any labeling of the points in A, that is $2^{|A|} = 2^{2d+1}$ possible labels (labeled by the given algorithm), we found one classifier to learn that labeling. So (3) is proven
$\Rightarrow VCdim(\mathcal{H}) \geq |A| = 2d + 1$ (9)

To prove (4) it is necessary that $\forall A, |A| = 2d + 2$, $\mathcal{H}$ does not shatter A. Let $A = \{A_i | A_i \in \mathbb{R}^2, i \in \overline{1, 2d+1}\}$. Here are 2 cases:

Case1: $\exists A_j \in A\ s.t. A_j \in CH(A_1, \dots, A_{j-1}, A_{j+1}, \dots, A_{2d+2})$. Then the label $(1,1,\dots,1,\underbrace{0}_{j},1,1,\dots,1)$ cannot be learned. For this we will use Reductio ad Absurdum Method: Suppose that: $\exists h_P \in \mathcal{H}$ such that $L_{A(h_A)} = 0$

$$\Rightarrow \exists a_{1_P}, a_{2_P}, \dots, a_{d_P}\ such\ that \begin{cases} A_i = 1, i \in \overline{1, 2d+1}/j \\ A_j = 0 \end{cases} \Rightarrow \begin{cases} A_i \in CH(a_{1_P}, a_{2_P}, \dots, a_{d_P}), i \in \overline{1, 2d+1}/j & (10) \\ A_j \notin CH(a_{1_P}, a_{2_P}, \dots, a_{d_P}) & (11) \end{cases}$$

$$\left. \begin{array}{c} (10) \\ A_j \in CH(A_1, \dots, A_{j-1}, A_{j+1}, \dots, A_{2d+2}) \end{array} \right| \Rightarrow \left. \begin{array}{c} A_j \in CH(a_{1_P}, a_{2_P}, \dots, a_{d_P}) \\ (11) \end{array} \right| \Rightarrow \perp \Rightarrow \nexists h_P \in \mathcal{H}\ \text{s.t.}\ L_{A(h_A)} = 0\ \ (12)$$

Case1: $\nexists A_j \in A\ s.t. A_j \in CH(A_1, \dots, A_{j-1}, A_{j+1}, \dots, A_{2d+2}) \Rightarrow (A_1, \dots, A_{2d+2})\ 2d + 2 - gon$ (13). Then the label $(1,0,1,0,\dots)$ cannot be learned. For this we will use Reductio ad Absurdum Method: Suppose that: $\exists h_P \in \mathcal{H}$ such that $L_{A(h_A)} = 0 \Rightarrow \exists a_{1_P}, a_{2_P}, \dots, a_{d_P}$ (14)

$$(13), (14), (A1), (A2) \Rightarrow \exists (m_i)_{i \in \overline{1, d+1}}\ m\ is\ the\ edge\ of\ the\ d - gon\ classifier \Rightarrow \\ \perp (d\ +\ 1\ edges\ from\ a\ d - gon) \Rightarrow \nexists h_P \in \mathcal{H}\ \text{s.t.}\ L_{A(h_A)} = 0\ (14)$$

$$(13), (14) \Rightarrow \nexists h_P \in \mathcal{H}\ \text{such that}\ L_{A(h_A)} = 0 \Rightarrow VCdim(\mathcal{H}) < |A| = 2d + 2\ (15)$$

$$(9), (15) \Rightarrow \boxed{VCdim(\mathcal{H}) = 2d + 1} \blacksquare$$