Advanced Machine Learning Assignment 1

Olteanu Fabian Cristian

FMI, AI Master, Year 1

---

1. Exercise 1

1.a.

Let $\mathcal{H}$ be our hypothesis class, where: $\mathcal{H} = \{h_{\omega_0,\omega_1,\omega_2,...,\omega_{2022}} : \mathbb{R}^{2022} \to \{0,1\} | h_{\omega_0,\omega_1,\omega_2,...,\omega_{2022}}(x) = \mathbf{1}_{\omega_1 x_1 + ... + \omega_{2022} x_{2022} \le \omega_0}(x), \omega_0, \omega_1, \omega_2, ..., \omega_{2022} \in \{1,2\}\}$.

In other words, $\mathcal{H}$ is a finite class (because components of ω and ω_0 can only be either 1 or 2, so $|\mathcal{H}| = 2^{2023}$) composed of classifiers which output positives depending whether the 2022-dimensional vector x belongs in some halfspace. Furthermore, if we denote $\omega = (\omega_1, ..., \omega_{2022}) \in \mathbb{R}^{2022}$, the following expressions are equivalent:

$$h_\omega(x) = \mathbf{1}_{\omega_1 x_1 + ... + \omega_n x_n \le \omega_0} = \mathbf{1}_{\omega \cdot x \le \omega_0}.$$

In order to justify the fact that $VCdim(\mathcal{H}) = 2023$, we first have to prove that $VCdim(\mathcal{H}) \ge 2023$ and afterwards $VCdim(\mathcal{H}) < 2024$. We can consider the standard basis of $\mathbb{R}^{2022}$ plus the origin ($e_0 = \mathbf{0}_{2022}$) as a set $B = \{e_0, ..., e_{2022}\}$ and prove that it is shattered by $\mathcal{H}$. Given a certain labelling $l_0, ..., l_{2022}$ to these points, we set the following relations:

$$\omega_0 = -l_0,$$
$$\omega_i = \omega_0 + l_i, i = \overline{1, 2022},$$

so $\omega \cdot e_0 - \omega_0 = l_0$ and for all $i = \overline{1, 2022}$, $\omega \cdot e_i - \omega_0 = l_i$ (for example, $\omega \cdot e_1 - \omega_0 = \omega_1 - \omega_0 = \omega_0 + l_1 - \omega_0 = l_1$. This proves that B is shattered by $\mathcal{H}$ and, since $|B| = 2023$, $VCdim(\mathcal{H}) \ge 2023$. Additionally, this statement holds for any $\omega_0, ..., \omega_{2022} \in \mathbb{R}$.

The proof of $VCdim(\mathcal{H}) < 2024$ can be achieved by utilizing Radon's Lemma, which states that for a set S from $\mathbb{R}^d$, $|S| = d+2$, there are two subsets of S with the property that their convex hulls intersect. Starting from this theorem, we can construct a set $S = \{x_1, ..., x_{2024}\} \subset \mathbb{R}^{2022}$ and assign to each element the labels $L = \{l_1, ..., l_{2024}\}$. Now, if we were to split S according to Radon's Lemma, we would arrive at the conclusion that one point always lies in the convex hulls of both subsets from S ($\mathcal{H}$ will never be able to realize the labels of S; the proof follows the demonstration written below in exercise 3 very closely), so $VCdim(\mathcal{H}) < 2024$ [1].

Thus, it is proven that $VCdim(\mathcal{H}) = 2023$ for any $\omega \in \mathbb{R}^{2022}$ and $\omega_0 \in \mathbb{R}$, so the original statement is also valid.

1.b.

Let $\mathcal{H}$ be our hypothesis class, where: $\mathcal{H} = \{h_{\omega_0,\omega_1,\omega_2,...,\omega_{2022}} : \mathbb{R}^{2022} \in \{0,1\} | h_{\omega_0,\omega_1,\omega_2,...,\omega_{2022}}(x) = \mathbf{1}_{\omega_1 x_1 + ... + \omega_{2022} x_{2022} \le \omega_0}(x), \omega_0, \omega_1, \omega_2, ..., \omega_{2022} \in \mathbb{R}\}$. $\mathcal{H}$ is infinite because $\omega_0, ..., \omega_{2022} \in \mathbb{R}$. $VCdim(\mathcal{H})$ is also 2023 as described in subsection a.

1.c.

Let $\mathcal{H} = \{h_\theta : [-1,1] \to \{0,1\} | h_\theta(x) = \mathbf{1}_{sin(\theta x) \geq 0}(x), \theta \in \mathbb{R}\}$. If we were to consider some set $X = \{x_1, ..., x_n\} \subset [-1,1]$, $\mathcal{H}$ can shatter X because $sin(\theta x)$ can oscillate at any frequency to accommodate labeling X. Hence, $VCdim(\mathcal{H}) = \infty$.

2.

We have $\mathcal{H} = \{h_a : \mathbb{R}^3 \to \{0,1\} | h_a(x) = \mathbf{1}_{[||x||_2 \leq a]}(x), x = (x_1, x_2, x_3) \in \mathbb{R}^3, ||x||_2 = \sqrt{x_1^2 + x_2^2 + x_3^2}\}$.

2.a.

$\mathcal{H}$ is PAC-learnable if there exists a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and there exists a learning algorithm A with the property that for every $\varepsilon, \delta > 0$, for every classifier $f \in \mathcal{H}$, for every distribution $\mathcal{D}$ on $\mathbb{R}^3$, when we run the learning algorithm A on a training set S consisting of $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ examples sampled independent and identically distributed from $\mathcal{D}$ and labeled by f, the algorithm A returns a hypothesis $h_S \in \mathcal{H}$ such that, with probability at least $1 - \delta$, the real risk of h_S is smaller than ε:

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

Let's consider the realizability assumption: there exists some $f = h_a^* \in \mathcal{H}$ such that $L(h_a^*) = 0$ ($a^* \in \mathbb{R}$), where $h_a^*(x) = \mathbf{1}_{[||x||_2 \leq a^*]}$. We construct a training set $S = \{(x_1, y_1), ..., (x_m, y_m) | y_i = h_a^*(x_i), x_i \in \mathbb{R}^3\}$. h_a^* labels each point from S positively if it is contained within the 3-d ball of radius a^* and labels negatively all other points (label 0).

Consider the following algorithm A, which takes as input the training set S and outputs $h_S = h_{aS}(x)$:

1. Take $a_S = \max\limits_{\substack{i=\overline{1,m} \\ y_i = 1}} (||x_i||_2)$ if there is at least a positively labeled sample in S (h_{a_S} is the ball of radius a_S), or
2. Take $a_S = -1$ if there is no positively labeled sample in S (h_{a_S} always outputs negatives).
3. Output $A(S) = h_{a_S}$.

By design, A is an ERM, so $L_{h^*,\mathcal{D}}(h_S) = 0$.

Consider $\mathcal{D}$ a distribution over $\mathcal{X} = \mathbb{R}^3$ and take $a_0 < a^* \in \mathbb{R}$ such that $\underset{x \sim \mathcal{D}_{\mathcal{X}}}{P}(||x||_2 \in (a_0, a^*) = \varepsilon$ (if $\mathcal{D}_{\mathcal{X}}(-\infty, a^*) \leq \varepsilon$ take $a_0 = -\infty$). Since $L_{\mathcal{D}}(h_S) > \varepsilon$ is equivalent with saying that $a_S < a_0$, we can say:

$$\underset{S \sim \mathcal{D}^m}{P}(L_{\mathcal{D}}(h_S) > \varepsilon) = P(a_S < a_0) = (1 - \varepsilon)^m \leq e^{-\varepsilon m} < \delta,$$

which, by definition, proves that $\mathcal{H}$ is PAC-learnable with the sample complexity $m \geq m_{\mathcal{H}}(\varepsilon, \delta) = \frac{1}{\varepsilon} \log \frac{1}{\delta}$.

2.b.

We know that $VCdim(\mathcal{H})$ is at least greater or equal than 1, since the hypothesis class contains non-constant classifiers, so it is able to shatter any set $S_0 = \{x_0 \in \mathbb{R}^3\}$, where $|S_0| = 1$, regardless of a.

Let's take $S = \{x_0, x_1\} \subset \mathbb{R}^3$, $|S| = 2$ and assign a labelling $L = \{l_0, l_1\}$. Our hypothesis class is composed of classifiers which output positives when the input is part of the origin-centered ball of radius a (B_a) and outputs negatives otherwise.

We need to prove that $\mathcal{H}$ can't shatter S, or in other words, to prove that all possible labellings L of the set S can't be realized by functions from $\mathcal{H}$. In our context there are four possible cases:

1. $x_0, x_1 \in B_a$,
2. $x_0 \in B_a$ and $x_1 \notin B_a$,
3. $x_0, x_1 \notin B_a$,
4. $x_0 \notin B_a$ and $x_1 \in B_a$.

The last two cases are equivalent to the first two if we change the labels, so we will only focus on the former two.

Case 1: It is obvious to see that there is no function in $\mathcal{H}$ that will correctly label points from S when the ground-truth is $L = \{0, 1\}$ or $L = \{1, 0\}$.

Case 2: Functions in $\mathcal{H}$ will mislabel points from S when $L = \{1, 1\}$ or $L = \{0, 0\}$.

Taking this into account, we can conclude that $\mathcal{H}$ does not shatter S, so $VCdim(\mathcal{H}) < 2$ and, furthermore, that $VCdim(\mathcal{H}) = 1$.

3.

$\mathcal{H} = \{h_{\theta_1,\theta_2} : \mathbb{R} \to \{0, 1\} | h_{\theta_1,\theta_2}(x) = h_{\theta_1,\theta_2}(x_1, x_2) = \mathbf{1}_{[\theta_1 + x_1 sin\theta_2 + x_2 sin\theta_2 > 0]}, \theta_1, \theta_2 \in \mathbb{R}\}$ is our hypothesis class. It can be observed that, similar to the example given in the first exercise, the class is also composed of classifiers which output positives depending whether the 2-dimensional vector x belongs in some halfspace. Following a proof very similar to the one described in the first exercise will render the following result: $VCdim(\mathcal{H}) = 3$.

First, we need to prove that there exists a set $C = \{c_0, c_1, c_2\} \subset \mathbb{R}^3$ which is shattered by $\mathcal{H}$. Let's fix $c_0 = (0, 0), c_1 = (1, 0), c_2 = (0, 1)$ and assign the labelling $L = \{l_0, l_1, l_2\}$ to C. We want to find $h_{\theta_1,\theta_2}(x)$ such that $h_{\theta_1,\theta_2}(c_i) = l_i$, $\forall l_i \in \{0, 1\}$, $i = \overline{0, 2}$.

We will first make the following notations: $\omega_0 = \theta_1$, $\omega_1 = sin\theta_2$, $\omega_2 = cos\theta_2$ and set:

$$\begin{aligned} &\omega_0 = l_0, \\ &\omega_i = -\omega_0 + l_i, i \in \{1, 2\}. \end{aligned}$$

It immediately follows that $h_{\theta_1,\theta_2}(c_0) = \mathbf{1}_{l_0>0}$, $h_{\theta_1,\theta_2}(c_1) = \mathbf{1}_{l_1>0}$ and $h_{\theta_1,\theta_2}(c_2) = \mathbf{1}_{l_2>0}$, which are always going to output the ground-truth labels, so $\mathcal{H}$ shatters C and $VCdim(\mathcal{H}) \geq 3$.

Proving that $VCdim(\mathcal{H}) < 4$ can be done following the same way of thinking as in the previous example. More explicitly, we start by defining our set $X = \{x_1, x_2, x_3, x_4\}$. If we consider the following system of equations:

$$\begin{pmatrix} x_1 & x_2 & x_3 & x_4 \\ 1 & 1 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{pmatrix} = 0,$$

with the variables $\lambda_1, \lambda_2, \lambda_3, \lambda_4$, we know that there exist solutions $\lambda_{0i} \neq 0, i \in \mathbb{N}, \lambda_{0i} \in \mathbb{R}$, because there are only two equations and four variables.

If we construct $P = \{i | \lambda_{0i} > 0\}$ and $N = \{j | \lambda_{0j} < 0\}$, we get the following relation:

$$\lambda^* = \sum_{i \in P} \lambda_{0i} = -\sum_{j \in N} \lambda_{0j} \neq 0.$$

Additionally, we know that $\displaystyle\sum_{i=1}^{4} \lambda_{0i} x_i = 0$, so we can write:

$$x^* = \sum_{i \in P} \lambda_{0i} x_i = -\sum_{j \in N} \lambda_{0j} x_j \neq 0.$$

Now, if we compute the point

$$\frac{x^*}{\lambda^*} = \sum_{i \in P} \frac{\lambda_{0i}}{\lambda^*} x_i = -\sum_{j \in N} \frac{\lambda_{0j}}{\lambda^*} x_j,$$

we can observe that the point $\frac{x^*}{\lambda^*}$ lies in both convex hulls of $X_1 = \{x_i | i \in P\}$ and $X_2 = \{x_j | j \in N\}$, so this proves that, for example, label $\{1,1,1,0\}$ is not realizable.

We have thus used Radon's Theorem to prove that $\mathcal{H}$ cannot shatter $X = \{x_1, x_2, x_3, x_4\}$ and, as such, $VCdim(\mathcal{H}) < 4$, so $VCdim(\mathcal{H}) = 3$.

4.

4.a.

α denotes the aspect of the rectangles classifiers in $\mathcal{H}_\alpha$. We have three cases:

- if $0 < \alpha < 1$, the rectangle's width is longer than its height,
- if $\alpha = 1$, we have a square classifier,
- if $\alpha > 1$, the rectangle's height is longer than its width.

Let $S = \{(x_1, y_1), ..., (x_m, y_m) | y_i = h_{a_1^*, b_1^*, a_2^*, b_2^*}, x_i \in \mathbb{R}^2\}$ be our training set, so $h_{a_1^*, b_1^*, a_2^*, b_2^*} = \mathbf{1}_{R^*}$. In order to choose an algorithm A that is an ERM, we have to take into consideration the fact that both h^* and the output of A, $h_S = h_{a_{1S}, b_{1S}, a_{2S}, b_{2S}}$, have the same parameter α, so we have to take into account the three cases mentioned above. Furthermore, for each case, we have four additional edge cases:

1. when the points are close to the bottom edge of the rectangle,
2. when the points are close to the leftmost edge of the rectangle,
3. when the points are close to the rightmost edge of the rectangle,
4. when the points are close to the top edge of the rectangle.

The algorithm will have to compensate the position of the rectangle of h_S such that it doesn't go outside R^*, because the aspect ratio of both h^* and h_S have to be the same.

5.

$\mathcal{H} = \{h_\theta : \mathbb{R} \to \{0,1\} | h_\theta(x) = \mathbf{1}_{[\theta, \theta+1] \cup [\theta+2, \theta+4] \cup [\theta+6, \theta+9]}(x), \theta \in \mathbb{R}\}$ is our hypothesis class. We will assume that $VCdim(\mathcal{H}) = 4$.

The first step is to prove that $VCdim(\mathcal{H}) \geq 4$. Consider the set $C_0 = \{c_1, c_2, c_3, c_4 | c_1 \leq c2 \leq c_3 \leq c_4\} \subset \mathbb{R}$ and the labelling $L = \{l_1, l_2, l_3, l_4\}$, $l_i \in \{0,1\}, i = \overline{1,4}$. Take $c_1 = -5$, $c_2 = -4.2$, $c_3 = -1.8$, $c_4 = 1.6$ and prove that all $2^4 = 16$ labels can be realized by classifiers from $\mathcal{H}$. In other words, find at least a value θ_i for each labelling such that h_{θ_i} correctly labels C:

1. For $L = \{0,0,0,0\}$, take $\theta_1 = -20.0$, $h_{\theta_1} = \mathbf{1}_{[-20.0,-19.0] \cup [-18.0,-16.0] \cup [-14.0,-11.0]}$
2. For $L = \{0,0,0,1\}$, take $\theta_2 = -6.1$, $h_{\theta_2} = \mathbf{1}_{[-6.1,-5.1] \cup [-4.1,-2.1] \cup [-0.1,2.9]}$
3. For $L = \{0,0,1,0\}$, take $\theta_3 = -10.1$, $h_{\theta_3} = \mathbf{1}_{[-10.1,-9.1] \cup [-8.1,-6.1] \cup [-4.1,-1.1]}$
4. For $L = \{0,0,1,1\}$, take $\theta_4 = -2.4$, $h_{\theta_4} = \mathbf{1}_{[-2.4,-1.4] \cup [-0.4,1.6] \cup [3.6,6.6]}$
5. For $L = \{0,1,0,0\}$, take $\theta_5 = -10.9$, $h_{\theta_5} = \mathbf{1}_{[-10.9,-9.9] \cup [-8.9,-6.9] \cup [-4.9,-1.9]}$
6. For $L = \{0,1,0,1\}$, take $\theta_6 = -6.9$, $h_{\theta_6} = \mathbf{1}_{[-6.9,-5.9] \cup [-4.9,-2.9] \cup [-0.9,2.1]}$

5

7. For $L = \{0,1,1,0\}$, take $\theta_7 = -10.8$, $h_{\theta_7} = \mathbf{1}_{[-10.8,-9.8]\cup[-8.8,-6.8]\cup[-4.8,-1.8]}$
8. For $L = \{0,1,1,1\}$, take $\theta_8 = -4.9$, $h_{\theta_8} = \mathbf{1}_{[-4.9,-3.9]\cup[-2.9,-0.9]\cup[1.1,4.1]}$
9. For $L = \{1,0,0,0\}$, take $\theta_9 = -14.0$, $h_{\theta_9} = \mathbf{1}_{[-14.0,-13.0]\cup[-12.0,-10.0]\cup[-8.0,-5.0]}$
10. For $L = \{1,0,0,1\}$, take $\theta_{10} = -6.0$, $h_{\theta_{10}} = \mathbf{1}_{[-6.0,-5.0]\cup[-4.0,-2.0]\cup[0.0,3.0]}$
11. For $L = \{1,0,1,0\}$, take $\theta_{11} = -9.0$, $h_{\theta_{11}} = \mathbf{1}_{[-9.0,-8.0]\cup[-7.0,-5.0]\cup[-3.0,0.0]}$
12. For $L = \{1,0,1,1\}$, take $\theta_{12} = -5.8$, $h_{\theta_{12}} = \mathbf{1}_{[-5.8,-4.8]\cup[-3.8,-1.8]\cup[0.2,3.2]}$
13. For $L = \{1,1,0,0\}$, take $\theta_{13} = -13.2$, $h_{\theta_{13}} = \mathbf{1}_{[-13.2,-12.2]\cup[-11.2,-9.2]\cup[-7.2,-4.2]}$
14. For $L = \{1,1,0,1\}$, take $\theta_{14} = -7.4$, $h_{\theta_{14}} = \mathbf{1}_{[-7.4,-6.4]\cup[-5.4,-3.4]\cup[-1.4,1.6]}$
15. For $L = \{1,1,1,0\}$, take $\theta_{15} = -5.2$, $h_{\theta_{15}} = \mathbf{1}_{[-5.2,-4.2]\cup[-3.2,-1.2]\cup[0.8,3.8]}$.
16. For $L = \{1,1,1,1\}$, take $\theta_{16} = -8.2$, $h_{\theta_{16}} = \mathbf{1}_{[-8.2,-7.2]\cup[-6.2,-4.2]\cup[-2.2,0.8]}$

The values were obtained by using a search algorithm implemented in Python, which tried θ_i values from a set $\{-20, -19.9, ..., 19.9, 20\}$ until one value satisfies the labelling of some points, which were also searched for via brute force. The conclusion is that $\mathcal{H}$ shatters C_0, so $VCdim(\mathcal{H}) \geq 5$.

Consider the set $C = \{c_1, c_2, c_3, c_4, c_5 | c_1 \leq c_2 \leq c_3 \leq c_4 \leq c_5\} \subset \mathbb{R}$ and the labels $L_1 = \{1,0,1,0,1\}$. Assuming that $\mathcal{H}$ shatters C, then there exists some $h_\theta \in \mathcal{H}$ which correctly labels the points from C. One possible configuration can be the following:

$$\begin{aligned}
\theta \leq c_1 &\leq \theta + 1 \\
\theta + 1 < c_2 &< \theta + 2 \\
\theta + 2 \leq c_3 &\leq \theta + 4 \\
\theta + 4 < c_4 &< \theta + 6 \\
\theta + 6 \leq c_5 &\leq \theta + 9,
\end{aligned}$$

which means:

$$\begin{aligned}
2 \leq c_3 - c_1 &\leq 3 && (1) \\
3 < c_4 - c_2 &< 4 && (2) \\
2 \leq c_5 - c_3 &\leq 5 && (3) \\
6 \leq c_5 - c_1 &\leq 8. && (4)
\end{aligned}$$

Consider the same configuration, but with the labels $\{1,0,0,1,1\}$. We will prove that, no matter how the points are positioned, there is no way this labelling can be achieved by any $h_\theta \in \mathcal{H}$. We have two possible ways of choosing the points:

<u>Case 1</u>: The first point is in the first interval, the next two are outside of it and the last two are in the second interval,

<u>Case 2</u>: The first point is in the second interval, the next two are outside of it and the last two are in the last interval.

The first case implies that $2 \leq c_5 - c_1 \leq 3$, but if we substract it from relation (4), we get $4 < 0 < 5$, which is impossible.

The second case implies that $4 \leq c_5 - c_1 \leq 5$, but if we substract if from (4), we get $2 \leq 0 \leq 3$, which is also impossible.

We've thus proven that any configuration from C cannot satisfy these two labels, so $\mathcal{H}$ doesn't shatter any C, which means that $VCdim(\mathcal{H}) < 5$, so $VCdim(\mathcal{H}) = 4$.

References

[1] Stefan Hausler, VC Dimension, Tutorial for the Course Computational Intelligence, https://www2.spsc.tugraz.at/www-archive/downloads/vc_examples.pdf