2019 ML foundation HW 2

B07902064 資工二 蔡銘軒

December 13, 2019

Problem 1

Problem 2

VC-dimension no less than 4 $\iff$ There exits a set of 4 points that can be shattered.
Consider the set of points $\{(1,0),(0,1),(-1,0),(0,-1)\}$, we illustrate all the dichotomies by discussing how many points the rectangle covers:
zero point and four points:

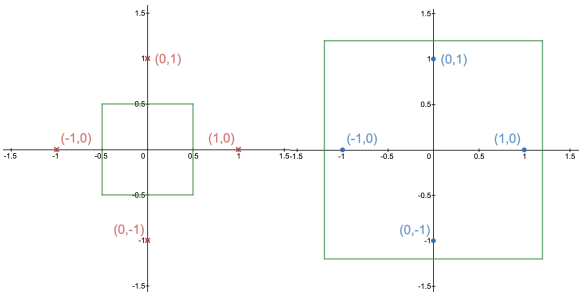


Figure 1: The 2 dichotomies are shown as above

One point:

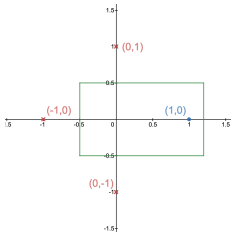


Figure 2: By symmetry, we can generate 4 dichotomies by covering one point

two points:

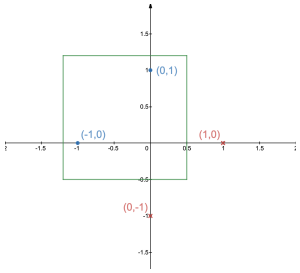


Figure 3: By symmetry, we can generate 4 dichotomies by covering two adjacent points

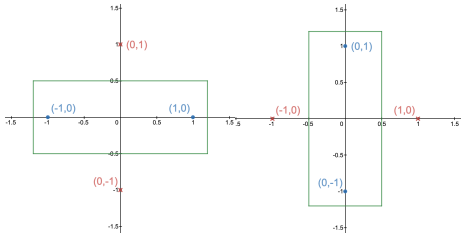


Figure 4: The rectangle can also cover two diagonal points

three points:

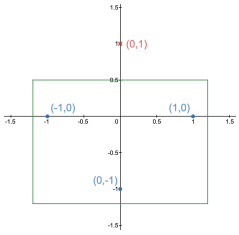


Figure 5: By symmetry, we can generate 4 dichotomies by covering three points

We have shown how to obtain all the $2^4 = 16$ dichotomies, and thus the VC-dimension is no less than 4.

Problem 3

The VC-dimension is ∞, as for every $n \in \mathbb{N}$, there is a set of n points that $\mathcal{H}$ can shatter. The proof is as follows:

To start with, we shift from the usual base 10 system to base 4 system and we will express α in base 4. For each $x = (4^i)_{10}$, $i \in \mathbb{N} \cup \{0\}$, αx can be viewed as left-shifting α by i digits. For example, $(33.22)_4 \cdot (4)_{10} = (33.22)_4 \cdot (10)_4 = (332.2)_4$.
The modulo (mod 4) is then the 4^0 and the decimal part of αx. For example, $(33.22)_4 \bmod 4 = (3.22)_4$
Based on the above idea, by letting $\text{sign}(0) = -1$, we have:

$$h_\alpha(x) = \begin{cases} 1 & \text{if the coefficient of } 4^0 \text{ in } \alpha x \text{ is } 0 \\ -1 & \text{if the coefficient of } 4^0 \text{ in } \alpha x \text{ is } 1 \end{cases}$$

Let $x_1, x_2, \ldots, x_n$ be the n points, such that $x_i = (4^i)_{10}$ for $1 \leq i \leq n$, then we can generate **every** dichotomy by manipulating α by letting $\alpha = (0.a_1a_2...a_n)_4$, such that

$$a_i = \begin{cases} 0 & \text{if we want to let } h_\alpha(x_i) = 1 \\ 1 & \text{if we want to let } h_\alpha(x_i) = -1 \end{cases} \quad \text{for } i = 1, 2, \ldots, n$$

The proof holds for every $n \in \mathbb{N}$, so the VC-dimension is ∞.

Problem 4

Proof by contradiction:

Let $d_{vc}(\mathcal{H}_1 \cap \mathcal{H}_2) = n$, $d_{vc}(\mathcal{H}_1) = m$ such that $n > m$.
It means that there exists a set $\mathcal{D}$ with n elements such that $\mathcal{H}_1 \cap \mathcal{H}_2$ can shatter, and $\mathcal{H}_1$ cannot. But $\mathcal{H}_1 \cap \mathcal{H}_2 \subseteq \mathcal{H}_1$, so the hypotheses in $\mathcal{H}_1 \cap \mathcal{H}_2$ that shatter $\mathcal{D}$ must also be in $\mathcal{H}_1$, which means $\mathcal{H}_1$ can also shatter $\mathcal{D}$ and thus $d_{vc}(\mathcal{H}_1)$ is no less than n, contradiction.

Problem 5

Consider N points on a line. There are $N+1$ intervals in which we can place the threshold θ. For each θ, we can decide whether it's positive ray or negative ray. However, we note that placing the threshold to the left of all points is equivalent to placing it to the right of all points, as they mark all the points as either all positive or all negative, so they generate the same dichotomies. So we only have $2N$ different hypotheses, and $m_{\mathcal{H}_1 \cup \mathcal{H}_2} = 2N$.
We note that when $N = 2$, $2N = 2^N = 4$, but when $N = 3$, $2N = 6 \neq 2^N = 8$. By definition, $d_{vc}(\mathcal{H}_1 \cup \mathcal{H}_2) = 2$.

Problem 6

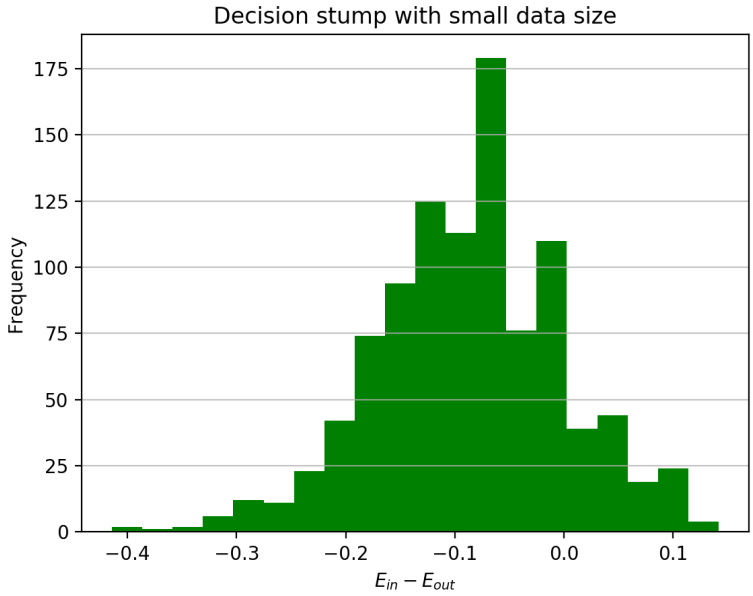
We first consider the case where we have a hypothesis h that makes an error with probability μ in approximating a deterministic function f. If we use the same h to approximate a noisy version of f that has a probability λ that f output correctly, and probability $1 - \lambda$ that the output of f is "flipped", then it has an error rate $\mu\lambda + (1 - \mu)(1 - \lambda)$.
The calculation is straightforward. h makes an error when $h(x) = f(x)$, but $f(x)$ is flipped. This gives $(1 - \mu)(1 - \lambda)$. h also makes an error when $h(x) \neq f(x)$, and $f(x)$ is not flipped. This gives $\mu\lambda$. Hence the error rate is $\mu\lambda + (1 - \mu)(1 - \lambda)$.
For a hypothesis $h_{s,\theta}$, let $s \in \{1, -1\}$, representing positive ray when $s = 1$ and negative ray otherwise. Its error rate μ is:

$$\mu = \begin{cases} \frac{|\theta|}{2} & \text{if } s = 1 \\ \frac{2-|\theta|}{2} & \text{if } s = -1 \end{cases}$$

which can be written as $\mu = \frac{s(|\theta|-1)+1}{2}$. Combined with the aforementioned property, let $\lambda = 0.8$, we have:

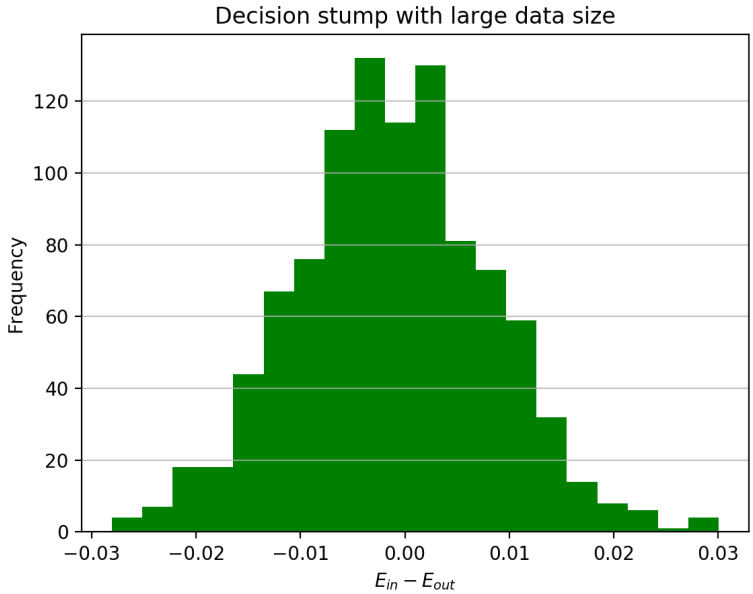
$$\begin{aligned} E_{out} &= \mu\lambda + (1 - \mu)(1 - \lambda) \\ &= \frac{s(|\theta| - 1) + 1}{2} \cdot \frac{4}{5} + \frac{1 - s(|\theta| - 1)}{2} \cdot \frac{1}{5} \\ &= 0.3s(|\theta| - 1) + 0.5 \end{aligned}$$

Problem 7



In most cases, E_{in} and E_{out} differs by around 0.1. In fact, the average of $E_{in} - E_{out}$ is around -0.1. The distribution is affected by the noise. Although the flip rate is 20%, with a small data size, it is likely that the noise rate does not approximate to 20% and cause E_{in} to deviate from expectation. This is reflected in the histogram that some E_{in} and E_{out} differs by more than 0.3, which is relatively large.

Problem 8



With a large data set, E_{in} and E_{out} is much closer. In most cases, they differs by no more than 0.01. Also, the histogram has a balanced distribution. When the data size is larger, the noise rate is more likely to be

close to 20%, so the outcome is more stable. Compared to the result in the previous problem, we find that the difference between E_{in} and E_{out} has shrinked significantly. This verifies the Hoeffding inequality and the VC bound that we can achieve better learning outcome by increasing the data size. Also, when the data size is larger, the noise rate is more accurate, so the distribution is more balanced than the previous one.

Bonus

We first show that $d_{vc}(\mathcal{H}) \geq 2^d$.
For given thresholds $\mathbf{t} = \{t_1, t_2, \ldots, t_d\}$, every vector $\mathbf{x} = [x_1, x_2, \ldots, x_d]^T \in \mathbb{R}^d$ is classified into one of the 2^d regions by the relation $x_i > t_i$ or $x_i \leq t_i$ for $i = 1, 2, \ldots, d$. We choose a set of data $\mathcal{D} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{2^d}\}$ such that for $i \neq j$, $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to different regions. Let $\mathbf{v}_i \in \{0, 1\}^d$ be the corresponding vector to $\mathbf{x}_i$ for $i = 1, 2, \ldots, 2^d$, then $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{2^d}\} = \{0, 1\}^d$. We can easily manipulate $h_{\mathbf{t},\mathbf{S}}(\mathbf{x}_i)$ by choosing whether $\mathbf{v}_i \in S$ for $i = 1, 2, \ldots, 2^d$. Since the choice for each $\mathbf{v}$ is independent, we can generate 2^{2^d} different sets S, which corresponds to 2^{2^d} different dichotomies under the chosen $\mathbf{t}$ and $\mathcal{D}$.

Next we show $d_{vc}(\mathcal{H}) < 2^d + 1$.
The proof is by induction on the number of thresholds. We will prove that with n thresholds in $\mathbb{R}^d$, where $1 \leq n \leq d$, any set of $2^n + 1$ points cannot be shattered.
Base case: When $n = 1$, the vectors in $\mathbb{R}^d$ are classified only by one threshold t_i for some $i \in [1, d]$. The space is divided into two regions. The hypothesis set is effectively equivalent to one dimensional PLA, and thus its VC dimension is 2. So it cannot shatter $2^1 + 1 = 3$ points.
Induction hypothesis: Assume with n thresholds, we cannot shatter any set of $2^n + 1$ points.
Induction step: Assume for the contradiction that with $n+1$ thresholds, we can shatter $2^{n+1} + 1$ points. Suppose the thresholds are $\{t_1, t_2, \ldots, t_{n+1}\}$, then the vectors in $\mathbb{R}^d$ can first be divided into two parts according to whether $x_i > t_i$ for some $i \in [1, n + 1]$. It is obvious that one part contains at least $2^n + 1$ points or more. As shattering $2^{n+1} + 1$ points means we can also shatter any subset of these points, so we can shatter $2^n + 1$ points that lie in the same part with the remaining n thresholds, which is a contradiction. So with $n + 1$ thresholds in $\mathbb{R}^d$, we cannot shatter any set of $2^{n+1} + 1$ points.
Following the proof, we cannot shatter $2^d + 1$ points with d thresholds in $\mathbb{R}^d$, so $d_{vc}(\mathcal{H}) < 2^d + 1$.

Now we have $d_{vc}(\mathcal{H}) \geq 2^d$ and $d_{vc}(\mathcal{H}) < 2^d + 1$, we conclude that $d_{vc}(\mathcal{H}) = 2^d$.