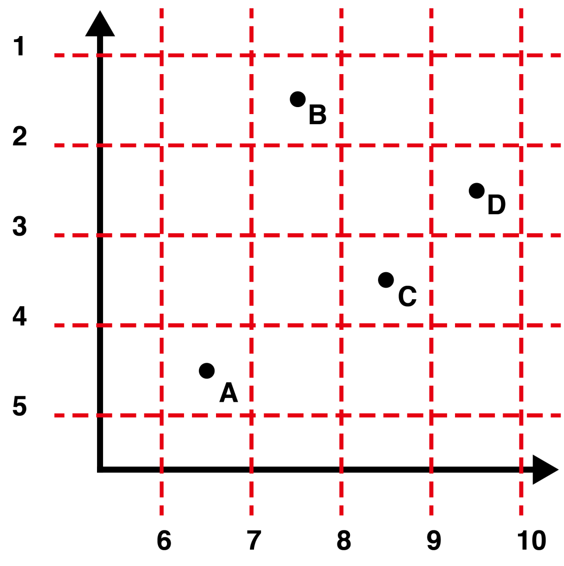
Machine Learning Foundations HW 2

R09946006 | 何青儒 | HO, Ching-Ru | Oct 29th, 2020

---

1. [**c**] $(1,1,3),(7,8,9),(15,16,17),(21,23,25)$
    - [**a**] $(7,8,9),(17,18,19),(27,28,29)$ are on the same line. Because they are collinear, the middle point can't be shattered with other two points.
    - [**b**] $(1,1,1),(7,8,9),(15,16,17),(21,23,25)$ are on the same plane with $8x_1-16x_2+8x_3=0$. Because they are coplanar, the diagonal points can't be shattered with others points.
    - [**b**] $(1,1,3),(7,8,9),(15,16,17),(21,23,25)$ are not on the same plane. They can be shattered in some method.
    - [**d**] $(1,3,5),(7,8,9),(15,16,17),(21,23,25)$ are on the same plane with $8x_1-16x_2+8x_3=0$. Because they are coplanar, the diagonal points can't be shattered with others points.
    - [**e**] $(1,2,3),(4,5,6),(7,8,9),(15,16,17),(21,23,25)$ are on the same plane with $8x_1-16x_2+8x_3=0$. Because they are coplanar, the diagonal points can't be shattered with others points.
2. [**d**] $4N-2$
    - 

For the horizontal line:
(up area for O / down area for X)

line 1: XXXX / OOOO
line 2: XOXX / OXOO
line 3: XOXO / OXOX
line 4: OXXX / XOOO
line 5: OOOO / XXXX

For the vertical line:
(left area for O / right area for X)

line 6: XXXX / OOOO
line 7: OXXX / XOOO
line 8: OOXX / XXOO
line 9: OOOX / XXXO
line 10: OOOO / XXXX

Notation: ABCD / reverse

    - If we have N points, we can draw $N+1$ lines to shatter each points, however, the first line and the last line will have the same label, e.g., if $N=4$ as above, the first line (line 1) is labeled as OOOO, and the last line (line 5) is labeled as XXXX, but XXXX is reverse of OOOO, so we actually have $2N$ labels, by line 1 to 4 (with reverse label). Then if we consider another axis, N points can draw $N+1$ lines, but due to the same reason, it still has $2N$ labels (line 6 and line 10 have the same result). After all, both two axis have the label of OOOO and XXXX, so it must minus these two condition (line 6 as line 1, and line 10 as line 5). Thus ,the growth function of axis-aligned perceptrons in 2D for $N \geq 4$ equals to $m_{\mathcal{H}}(N) = 2N + 2N - 2 = 4N - 2$.

3. [**d**] 3
    - 2D perceptrons with $w_0 > 0$ is just linear biased, it doesn't effect the **property** of 2D perceptrons without bias. Thus, the VC dimension of 2D positively-biased perceptrons is the same as without-biased perceptrons.
4. [**b**] $\binom{N+1}{2} + 1$
    - In fact, the range of $h(\mathbf{x})$ is a small sphere with radius a inside a big sphere with radius b (such as a small ball in a big ball, forming a hollow sphere), so if we project it on a surface, it will like a concentric circles with radius a and b ($b > a$). Between both circle is the range of $+1$, and otherwise is the range of -1.
    - The range is equivalent as the positive interval (one way, given 2 points and classify by three parts) mentioned in lecture slide. So the growth function of the hypothesis set is the same as positive interval, which means $m_{\mathcal{H}} = \binom{N+1}{2} + 1$.
5. [**b**] 2
    - According to the lecture slide, the VC dimension of positive interval is equal to 2.
6. [**d**] $2\sqrt{\frac{8}{N}\ln(\frac{4m_{\mathcal{H}}(2N)}{\delta})}$
    - From $\delta = \mathbb{P}[\exists h \in \mathcal{H} \text{ s.t.} |E_{in}(h) = E_{out}(h)| > \epsilon] \leq 4m_{\mathcal{H}}(2N)\exp(-\frac{1}{8}\epsilon^2 N)$, we know $|E_{out}(h) - E_{in}(h)| > \epsilon \geq \sqrt{\frac{8}{N}\ln(\frac{4m_{\mathcal{H}}(2N)}{\delta})}$ from the inequality. Separates LHS into three parts:
    - $E_{out}(g) - E_{out}(g*) = [E_{out}(g) - E_{in}(g)] + [E_{in}(g) - E_{in}(g*)] + [E_{in}(g*) - E_{out}(g*)]$
    - Due to $g = \text{argmin}_{h\in\mathcal{H}} E_{in}(h)$, we can find that $E_{in}(g)$ (the minimum $E_{in}(\cdot)$ ever) should smaller than $E_{in}(g*)$, making the middle term less than 0.
    - Because of $h,\ g,\ g* \in \mathcal{H}$, we can use the inequality as above. Thus, we can get
    $E_{out}(g) - E_{out}(g*) \leq [E_{out}(g) - E_{in}(g)] + [E_{in}(g*) - E_{out}(g*)]$
    $\leq |[E_{in}(g) - E_{out}(g)]| + |[E_{in}(g*) - E_{out}(g*)]|$
    $\leq 2 \cdot |E_{out}(h) - E_{in}(h)|$
    $\leq 2\sqrt{\frac{8}{N}\ln(\frac{4m_{\mathcal{H}}(2N)}{\delta})}$
7. [**d**] $\lfloor \log_2 M \rfloor$
    - If $\mathcal{H}$ shatters any set S in to $\{-1, +1\}$ (dichotomy), then $|\mathcal{H}| = M$ is at least 2^m, meaning the VC dimension can be at most $\log_2 \mathcal{H} = \log_2 M$. We choose $\lfloor \log_2 M \rfloor$ as the largest value in this case.
8. [**d**] $k + 1$
    - Because the function is symmetric, coordinates in k will be the monomials $\varnothing, k_1, \cdots, k_n, k_1k_2, \cdots, k_{n-1}k_n, \cdots, k_1k_2, \cdots, k_n$, Here the variable k_i indicates a 1 in the $i-th$ location of a binary n-vector. We divide the coordinates into $n+1$ classes $S_0, S_1, \cdots, S_n$ so that each class consists of all monomials of the same degree (matching the class index). Then a symmetric function h has the same value on all monomials in each class S_i. There are therefore $n+1$ degrees of freedom of functions in $\{-1, +1\}^k$.
    - For we can choose symmetric functions which evaluate independently on each of our $n+1$ classes Si of monomials. Hence we see that $\{-1, +1\}^k$ shatters a set S of $n+1$ coordinates, so long as there is one coordinate from each class S_i in S. On the other hand, it is also easy to see that there is no shattering of an $(n+2)$-set. For if we choose any collection of $n+2$ coordinates, then two of them have to be in the same class S_i. Hence every element of F does not distinguish these two coordinates, so shattering does not occur. This establishes that the VC dimension of the set of all the symmetric boolean functions exactly $n+1$.
    - Reference: Rubinstein, J. H., Rubinstein, B. I., & Bartlett, P. L. (2015). Bounding embeddings of VC classes into maximum classes. In *Measures of complexity* (pp. 303-325). Springer,

Cham.

9. [**c**] 3

   - By definition, when the inputs of set (N) is less than $d_{VC}(\mathcal{H}) = d$, means that it can be shattered in some situation, but not all of condition will. However, if the number of cases becomes $d+1$, every situation can not be shattered.
   - Thus, the first condition (some set of d distinct inputs is shattered by $\mathcal{H}$), the sixth condition (**some** set of $d+1$ distinct inputs is not shattered by $\mathcal{H}$) and the eighth condition (**any** set of $d+1$ distinct inputs is not shattered by $\mathcal{H}$) are correct.

10. [**c**] the sine family: the infinite number of hypotheses $\{h_\alpha : h_\alpha(x) = \text{sign}(\sin(\alpha \cdot x))\}$, for $x \in \mathbb{R}$

   - For all n, the set $S = \{2^1, 2^2, \cdots, 2^n\}$ is shattered by h. To see this, let $\alpha = -\pi \times (0.y_1 y_2 \cdots y_m)$ be a decimal binary encoding of a set of desired labels, converting -1 to 0. Essentially each x_i bit shifts α to produce the desired label as a result of the fact that $\text{sign}(\sin(\pi z)) = (-1)^{\lfloor z \rfloor}$. Thus, the VC dimension of this hypothesis class is infinite.

11. [**d**] $E_{out}(h, 0) = \frac{E_{out}(h,\tau)-\tau}{1-2\tau}$

   - It has probability τ to flip the output (having noise), which means that it also has probability $(1-\tau)$ doing nothing (having no noise). So the new $E_{out}(h, \tau)$ will equal to $[(\text{having noise}) \times (\text{having classification error in origin})]$, meaning it causes an error finally, and plus $[(\text{having no noise}) \times (\text{classify correct in origin})]$, meaning it causes an error finally.
   - Thus, $E_{out}(h, \tau) = (1-\tau)E_{out}(h, 0) + \tau(1 - E_{out}(h, 0))$, we can get $E_{out}(h, 0) = \frac{E_{out}(h,\tau)-\tau}{1-2\tau}$.

12. [**b**] 0.6

   - Due to $f(\mathbf{x}) = \text{argmax}_{i=1,2,3} x_i$, and $\mathbf{x}$ generates by a uniform $P(\mathbf{x})$ within $[0, 1]^3$, $P(f(\mathbf{x}) = 1) = P(f(\mathbf{x}) = 2) = P(f(\mathbf{x}) = 3) = \frac{1}{3}$.
   - Condition when $f(\mathbf{x}) = 1$, $y = \begin{cases} 1, \text{ with } P(y|\mathbf{x}) = 0.7, \text{ square error} = (1-1)^2 \times 0.7 = 0 \\ 2, \text{ with } P(y|\mathbf{x}) = 0.1, \text{ square error} = (1-2)^2 \times 0.1 = 0.1 \\ 3, \text{ with } P(y|\mathbf{x}) = 0.2, \text{ square error} = (1-3)^2 \times 0.2 = 0.8 \end{cases}$.
   Probability equals to $\frac{1}{3} \times (0 + 0.1 + 0.8) = 0.3$
   - Condition when $f(\mathbf{x}) = 2$, $y = \begin{cases} 1, \text{ with } P(y|\mathbf{x}) = 0.2, \text{ square error} = (2-1)^2 \times 0.2 = 0.2 \\ 2, \text{ with } P(y|\mathbf{x}) = 0.7, \text{ square error} = (2-2)^2 \times 0.7 = 0 \\ 3, \text{ with } P(y|\mathbf{x}) = 0.1, \text{ square error} = (2-3)^2 \times 0.1 = 0.1 \end{cases}$.
   Probability equals to $\frac{1}{3} \times (0.2 + 0 + 0.1) = 0.1$
   - Condition when $f(\mathbf{x}) = 3$, $y = \begin{cases} 1, \text{ with } P(y|\mathbf{x}) = 0.1, \text{ square error} = (3-1)^2 \times 0.1 = 0.4 \\ 2, \text{ with } P(y|\mathbf{x}) = 0.2, \text{ square error} = (3-2)^2 \times 0.2 = 0.2 \\ 3, \text{ with } P(y|\mathbf{x}) = 0.7, \text{ square error} = (3-3)^2 \times 0.7 = 0 \end{cases}$.
   Probability equals to $\frac{1}{3} \times (0.4 + 0.2 + 0) = 0.2$
   - Thus, $E_{out}(f) = 0.3 + 0.1 + 0.2 = 0.6$.

13. [**b**] 0.14

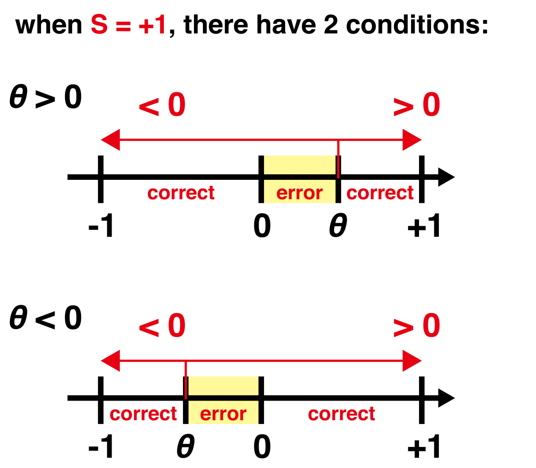
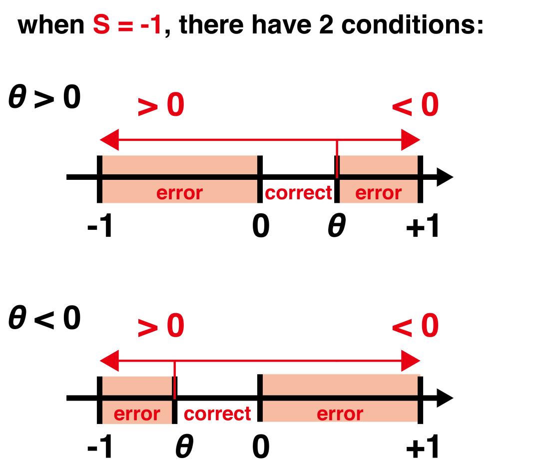
   - When $f(\mathbf{x}) = 1$, $f_*(\mathbf{x}) = 1 \times 0.7 + 2 \times 0.1 + 3 \times 0.2 = 1.5$, $f(\mathbf{x} - f_*(\mathbf{x})^2) = (1.5 - 1)^2 = 0.25$
   - When $f(\mathbf{x}) = 2$, $f_*(\mathbf{x}) = 2 \times 0.7 + 3 \times 0.1 + 1 \times 0.2 = 1.9$. $f(\mathbf{x} - f_*(\mathbf{x})^2) = (1.9 - 2)^2 = 0.01$
   - When $f(\mathbf{x}) = 3$, $f_*(\mathbf{x}) = 3 \times 0.7 + 1 \times 0.1 + 2 \times 0.2 = 2.4$. $f(\mathbf{x} - f_*(\mathbf{x})^2) = (2.4 - 3)^2 = 0.36$
   - Thus, $\Delta(f, f_*) = \mathbb{E}_{\mathbf{x} \sim P(\mathbf{x})}(f(\mathbf{x} - f_*(\mathbf{x}))^2 = \frac{1}{3} \times 0.25 + \frac{1}{2} \times 0.01 + \frac{1}{3} \times 0.36 = \frac{1}{3} \times 0.52 = 0.14$

14. [**d**] 12000

   - The growth function of decision stump is $m_{\mathcal{H}}(N) = 2N$, meaning $m_{\mathcal{H}}(2N) = 4N$. From the VC bound as given in the beginning of problem 6, we get $\delta \leq 4m_{\mathcal{H}}(2N)\exp(-\frac{1}{8}\epsilon^2 N) = 4 \cdot 4N \cdot \exp(-\frac{1}{8}\epsilon^2 N)$. From $\epsilon \leq \sqrt{\frac{8}{N}\ln(\frac{4m_{\mathcal{H}}(2N)}{\delta})}$:
   - [**a**]: $N = 6000$, $\epsilon = 0.1355221436442228$
   - [**b**]: $N = 8000$, $\epsilon = 0.11858486680810415$
   - [**c**]: $N = 10000$, $\epsilon = 0.10690374806230139$
   - [**d**]: $N = 12000$, $\epsilon = 0.09821010044458756$
   - [**e**]: $N = 14000$, $\epsilon = 0.09140799097244093$

- ϵ need to smaller than 0.1, and choose the smallest N, thus, we choose $N = 12000$ in this problem.

15. [**b**] $\frac{|\theta|}{2}$
    - $E_{out}(h_{+1,\theta}, 0)$ means there are no noise, and the direction indicator $s = +1$.
    - When $s = +1$, if $x \leq \theta, h(x) = -1$, else $x > \theta$, $h(x) = +1$.
    - $E_{out}(h_{+1,\theta}, 0) = \mathbb{P}(y = f(x) = +1, h_{s,\theta}(x) = -1) + \mathbb{P}(y = f(x) = -1, h_{s,\theta}(x) = +1)$
      $= \frac{1}{2} \times \frac{\theta}{1-(-1)}(\text{Yellow Area while } \theta > 0) + \frac{1}{2} \times \frac{|-\theta|}{1-(-1)}(\text{Yellow Area while } \theta < 0)$
      $= \frac{|\theta|}{2}$
    - By the same method used before, we can find E_{out} when $s = -1$:
    - $E_{out}(h_{-1,\theta}, 0) =$
      $\frac{1}{2} \times \frac{2-\theta}{1-(-1)}(\text{Orange Area while } \theta > 0) + \frac{1}{2} \times \frac{2-|-\theta|}{1-(-1)}(\text{Orange Area while } \theta < 0)$
      $= \frac{2-|\theta|}{2}$
    - Consider the probability τ, I'll use these E_{out} with noise function in the program:
        - $\begin{cases} s = +1 : E_{out}(h_{+1,\tau}) = (1-\tau)(\frac{|\theta|}{2}) + \tau(1 - \frac{|\theta|}{2}) = (1-\tau)(\frac{|\theta|}{2}) + \tau(\frac{2-|\theta|}{2}) \\ s = -1 : E_{out}(h_{-1,\tau}) = (1-\tau)(\frac{2-|\theta|}{2}) + \tau(1 - \frac{2-|\theta|}{2}) = (1-\tau)(\frac{2-|\theta|}{2}) + \tau(\frac{|\theta|}{2}) \end{cases}$.

<table>
<tr><th>$s = +1$</th><th>$s = -1$</th></tr>
<tr><td>
</td><td>
</td></tr>
</table>

16. [**d**] 0.30
17. [**b**] 0.02
18. [**e**] 0.40
19. [**c**] 0.05
20. [**a**] 0.00