Machine Learning Assignment 2

B08611035

October 2020

1 Perceptrons

(1) **c.**

The set of $\mathbf{x} \in \mathbb{R}^3$ can be shattered if the $\mathbf{x}$ is invertible (linear independent and determinant not equal to 0) such that $\text{sign}(\mathbf{w^T x}) = y$. In the case **a** and **e** the elements is linear dependent. and in the case **b** and **d**:

$$\begin{vmatrix} 1 & 1 & 1 & 1 \\ 1 & 7 & 8 & 9 \\ 1 & 15 & 16 & 17 \\ 1 & 21 & 23 & 25 \end{vmatrix} = 0$$

The determinant is zero, thus can not be shattered by the perceptron.

(2) **d.** $4N - 2$

Consider that 4 points of input arranging on a circle on the two dimensional coordinate without collinear and concurrent. Due to $w_1 w_2 = 0$, we can consider in only horizontal or vertical line to x-axis and y-axis respectively.

Consider the vertical ones first, the dichotomies of vertical line could be:

$$\begin{cases} (o, o, o, o) \\ (o, x, x, x) \\ (o, o, x, x) \\ (o, o, x, o) \\ (o, o, o, o) \end{cases}$$

So the combination will be (deduct the repeated point):

$$2\binom{N+1}{1} - 1 = 2N$$

And the results are the same in the horizontal case, so the total number of combination will be $4N$. And due to we have both (o, o, o, o) in both vertical and horizontal case, so the total number should minus 2 and will became $4N - 2$.

(3) **d.** 3

$\mathbf{w_0} > 0$ implies that the line does not pass through the origin. And we use same method in problem 2 above, we could not find a line such that (o, x, o, x). We still have infinite options of lines even we do not restrict $\mathbf{w_0}$. Thus, the VC dimension is 3 that is the same as two dimensional perceptron learning algorithm.

2 Ring Hypothesis Set

(4) **b.** $\binom{n+1}{2} + 1$

In three dimensional space we can use a ring hypothesis set, and actually it's hollow sphere that origin at $(0, 0, 0)$. We project it into two dimensional space (x-axis and y-axis) and a fixed z-axis in the same high for simplifying.

Consider that the inputs data number is $N = 4$, then we have interval between these 4 data with 1 to 5. Thus, we have:

$$\binom{N+1}{2}$$

combination, and we still have one more combination that all the points are -1 and make the it in the same interval. so the result combination will be:

$$\binom{N+1}{2} + 1$$

(5) **b.** 2

using the growth function in problem 4 above, as the inputs number $N = 2$, the combination for the hypothesis set is:

$$\binom{N+1}{2} + 1 = \binom{3}{2} + 1 = 3 + 1 = 4$$

And when the inputs number is $N = 3$, then the combination became:

$$\binom{N+1}{2} + 1 = \binom{4}{2} + 1 = 6 + 1 = 7 \leq 2^3$$

From the expression shown above, we can find that when the input number is $N = 3$, the dataset could not be shattered by the hypothesis, so that the break point is 3, and the VC dimension will be $k - 1$. So the VC dimension is 2.

3 Deviation from Optimal Hypothesis

(6) **d.** $2\sqrt{\frac{8}{N}\ln(\frac{4m_{\mathcal{H}}(2N)}{8})}$

We could first deep into the upper bound relationship

$$E_{out}(g) - E_{out}(g_*) = E_{out}(g) - E_{in}(g) + E_{in}(g) - E_{in}(g_*) + E_{in}(g_*) - E_{out}(g_*)$$

$$\leq E_{out}(g) - E_{in}(g) + E_{in}(g_*) - E_{out}(g_*) \leq \epsilon + \epsilon = 2\epsilon$$

Because the VC bound holds for any $g \in M$ and $|E_{out}(g) - E_{in}(g) \leq \epsilon$, ϵ is the upper bound. And due to:

$$P_{\mathcal{D}}\Big[|E_{out}(g) - E_{in}(g) \leq \epsilon\Big] \leq 4m_{\mathcal{H}}(2N)\exp(-\frac{1}{8}\epsilon^2 N)$$

for any $g = \mathcal{A}(\mathcal{D}) \in \mathcal{H}$. So that set the inference below,

$$\delta = 4m_{\mathcal{H}}(2N)\exp(-\frac{1}{8}\epsilon^2 N) \Rightarrow -\frac{1}{8}\epsilon^2 N = \ln(\frac{\delta}{4m_{\mathcal{H}}(2N)})$$

$$\Rightarrow \epsilon^2 = \frac{8}{N}\ln(\frac{4m_{\mathcal{H}}(2N)}{\delta}) \Rightarrow \epsilon = \sqrt{\frac{8}{N}\ln(\frac{4m_{\mathcal{H}}(2N)}{8})}$$

and Thus,

$$E_{out}(g) - E_{out}(g_*) \leq 2\epsilon = 2\sqrt{\frac{8}{N}\ln(\frac{4m_{\mathcal{H}}(2N)}{8})}$$

4 The VC Dimension

(7) **d.** $\lfloor \log_2 M \rfloor$

The hypothesis set $\mathcal{H}$ has hypothesis h for binary classification. So that:

$$h(x_1, x_2, ..., x_N) = (h(x_1), h(x_2,), ..., h(x_N)) \in \{\text{o,x}\}^N = 2^N$$

if $2^N > M$, means that $\mathcal{X}$ can not be shattered. Because the most distinct combination from hypothesis line is M, but we have 2^N combination. and if $2^N \leq M$ means the hypothesis line or hyperplane is equal or more than the combination of N inputs.

$$M \leq 2^N \Rightarrow N\log 2 \geq \log M$$

$$N \geq \frac{\log M}{\log 2} = \log_2 M$$

Notes that VC dimension of hypothesis set $d_{vc}(\mathcal{H})$ is the largest N for which $m_{\mathcal{H}}(N) = 2^N$, so that $d_{vc}(\mathcal{H}) = N$.

$$d_{vc}(\mathcal{H}) = N \geq log_2 M$$

Thus, largest possible value of $d_{vc}(\mathcal{H})$ is $\lfloor \log_2 M \rfloor$ for the answer.

(8) **d.** $k+1$

A symmetric boolean function:

$$h : \{-1, +1\}^k \rightarrow \{-1, +1\}$$

2^k combinations of $h(x)$ into y with binary classification:

$$2^k \times 2 = 2^{k+1}$$

(9) **c. 3**

Because $d_{vc}(\mathcal{H} = d$, it means that exists d inputs that could be shattered, and for the growth function is $m_{\mathcal{H}}(d) = 2^d$ which growth function take maximum of all possible $\mathcal{X} = (x_1, x_2, ..., x_N)$.

Thus, it means that there are set of d different inputs from the same distribution can be shattered by hypothesis set $\mathcal{H}$. And the $d+1$ is the break point, which means any set of $d+1$ inputs is not shattered by $\mathcal{H}$ that contain some of set with $d+1$ inputs can not be shattered by $\mathcal{H}$.

(10) **c.**

$$\Big\{ h_\alpha : h_\alpha(\mathbf{x}) = \text{sign}(\sin(\alpha \cdot \mathbf{x})) \Big\}$$

Assume that we have n inputs and output $(y_1, y_2, ..., y_n) \in [+1, -1]^n$. Let:

$$[x_1 = \frac{1}{2}, x_2 = \frac{1}{2^2} = \frac{1}{4}, ..., x_n = \frac{1}{2^n}]$$

if $y = 1$, $\sin(\alpha \cdot \mathbf{x}) > 0 \Rightarrow 0 < \alpha \cdot \mathbf{x} < \pi, ...$, and because $x \neq 0$, thus $0 < \mathbf{x} < \frac{\pi}{\alpha}, \frac{2\pi}{\alpha} < \mathbf{x} < \frac{4\pi}{\alpha}$. And if $y = -1$, $\sin(\alpha \cdot \mathbf{x}) < 0 \Rightarrow \pi < \alpha \cdot \mathbf{x} < 2\pi, 3\pi < \alpha \cdot \mathbf{x} < 4\pi$, and due to $x \neq 0$, thus $\frac{\pi}{\alpha} < \mathbf{x} < \frac{2\pi}{\alpha}, \frac{3\pi}{\alpha} < \mathbf{x} < \frac{4\pi}{\alpha}$.

In fact, $\alpha = \pi \mathbf{x}$ some constant that dependent to the results y_i and inputs x_i. Thus, set

$$\alpha = \pi \Big(1 + \sum_{i=1}^{n} 2^{i-1}(1 - y_i) \Big)$$

Then we could obtain the desired results of output.

5 Noise and Error

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

Define p as the correct number and n as the wrong number for the inference below:

$$\begin{aligned} E_{out}(h,\tau) &= \frac{p\cdot\tau + n(1-\tau)}{p+n} = \frac{n+\tau(p-n)}{p+n} \\ &= \frac{n}{p+n} + \frac{p-n}{p+n}\tau \end{aligned}$$

Note here that $p/(p+n) = E_{out}(h, 0)$.

$$E_{out}(h,\tau) = E_{out}(h,0) + \frac{p-n}{p+n}\tau = E_{out}(h,0) + \tau - 2E_{out}(h,0)\tau$$

$$\Rightarrow E_{out}(h,\tau) - \tau = E_{out}(h,0)(1-2\tau) \Rightarrow E_{out}(h,0) = E_{out}(h,0) = \frac{E_{out}(h,\tau)}{(1-2\tau)}$$

(12) **b.** 0.6

The distribution is shown as below:

$$P(y|\mathbf{x}) = \begin{cases} 0.7 & y = f(\mathbf{x}) \\ 0.1 & y = f(\mathbf{x})\text{mod } 3+1 \\ 0.2 & y = f(\mathbf{x}+1)\text{mode } 3+1 \end{cases}$$

so from the distribution above, we can infer the error function as below:

$$f(x) = \begin{cases} 1 & \text{err} = 0.1 + 0.2 \times 4 = 0.9 \\ 2 & \text{err} = 0.1 + 0.2 = 0.3 \\ 3 & \text{err} = 0.1 \times 4 + 0.2 = 0.6 \end{cases}$$

Thus,

$$\text{err} = P(y|\mathbf{x}) \cdot (f(x) - x)^2$$

$$E_{out}(f(X)) = \frac{0.9 + 0.6 + 0.3}{3} = 0.6$$

(13) **b.** 0.4

$$f_*(\mathbf{x}) = \sum_{y=1}^{3} y \cdot P(y|\mathbf{x})$$

and according to the description, the squared difference between f and f_* is:

$$\Delta(f, f_*) = \mathbb{E}_{\mathbf{x}\sim P(\mathbf{x})}(f(\mathbf{x}) - f_*(\mathbf{x}))^2$$

Thus, when $f(\mathbf{x})$ is equal to $1, 2$ and 3 respectively, the denoted target function is:

$$f_*(\mathbf{x}) = 1 \times (0.7) + 2 \times (0.1) + 3 \times (0.2) = 1.5$$

$$f_*(\mathbf{x}) = 2 \times (0.7) + 3 \times (0.1) + 1 \times (0.2) = 1.9$$

$$f_*(\mathbf{x}) = 3 \times (0.7) + 1 \times (0.1) + 2 \times (0.2) = 2.6$$

So the squared difference between f and f_* will be like below shown as the answers:

$$\Delta(f, f_*) = \frac{1}{3}\Big[(1 - 1.5)^2 + (2 - 1.9)^2 + (3 - 2.6)^2\Big] = 0.14$$

6 Decision Stump

(14) **d. 12000**

$$4m_{\mathcal{H}}(2N)\exp(-\frac{1}{8}\epsilon^2 N) \leq \delta$$

$$\Rightarrow 4(4N)\exp(-\frac{1}{8}\epsilon^2 N) = 16N\exp(-\frac{1}{8}\epsilon^2 N) \leq \delta$$

For the option **(a)** to **(c)**, the results values is not less than δ (0.1). And start from case **(d)** that come with the value less than δ.

$$16(12000)\exp(-\frac{1}{8}\epsilon^2(12000)) = 0.05873 \leq \delta = 0.1$$

And also for the case **e**, the outcome value is less than δ. as well. But pick the smaller N as the answer, the option d with $N = 12000$ will be a better choice.

$$16(14000)\exp(-\frac{1}{8}\epsilon^2(14000)) = 0.00563 \leq \delta = 0.1$$

(15) **b.** $1/2|\theta|$

Because $f(x) = \text{sign}(x)$, $h_{+1,\theta}(x) = \text{sign}(x - \theta)$. Thus the hypothesis is positive ray like.

$$h_{+1,\theta}(x)\begin{cases} = f(x) & \text{for } x \leq \theta \text{ and } x \geq 0 \\ \neq f(x) & \text{for } \theta \leq x \leq 0 \end{cases}$$

Due to the whole interval is $1 - (-1) = 2$, so the out-of-sample error will be:

$$E_{out}(h_{+1,\theta}m0) = \frac{1}{2}|\theta|$$

7 Programming*

(16) **d. 0.3**

(17) **b. 0.02**

(18) **e. 0.4**

(19) **c. 0.05**

(20) **a. 0.00**