Machine learning HW 3

Understanding Machine Learning: From Theory to Algorithms[Exercises for chapters 6 and 9]

Mehrnaz jalili:400422061- Arash Sajjadi:400422096

Email: Mehrnazjalili1991@gmail.com

Email: arash.sajjadi@yahoo.com

Date: December 14, 2021

chapter 6

Exercise 6.2:

The basic assumptions of the question: $|\mathcal{X}| < \infty \qquad k \leq |\mathcal{X}|$

1. $\mathcal{H}_{=k}^{\mathcal{X}} = \left\{ h \in \{0,1\}^{\mathcal{X}} : |\{x : h(x) = 1\}| = k \right\}$

   Let $C \subseteq \mathcal{X}$ such that $|C| = k+1 \Rightarrow \nexists h \in \mathcal{H}_{=k}$ s.t $h_C(x) = 1$ ①

   Let $C \subseteq \mathcal{X}$ such that $|C| = |\mathcal{X}| - k + 1 \Rightarrow \nexists h \in \mathcal{H}_{=k}$ s.t $h_C(x) = 0$ ②

   From ① and also ② we have $\text{VCdim}(\mathcal{H}_{=k}) \leq \min\{k, |\mathcal{X}| - k\}$

   Now if we prove that $\text{VCdim}(\mathcal{H}_{=k}) \geq \min\{k, |\mathcal{X}| - k\}$, the desired result will be achieved. So we will continue the proof path to $\text{VCdim}(\mathcal{H}_{=k}) \geq \min\{k, |\mathcal{X}| - k\}$

   Let $C = \{x_1, ..., x_m\}$ such that $m \leq min\{k, |\mathcal{X}| - k\}$. Get the label $(y_1, ..., y_m) \in \{0,1\}^m$ corresponds to C. also we donate $\sum y_i$ by s Let E be a set of $k - s$ arbitrary members of $\mathcal{X} - C$ so $E \subseteq \mathcal{X} - C$ ③

   suppose $h \in \mathcal{H}_{=k}$ be a hypothesis which satisfies $\forall X_i \in C : h(x_i) = y_i$ and, by the same function, we assign one member of label E to all members. Therefore, we were able to generate all possible functions on C with this set of hypotheses. ④

   From ③ and ④ we conclude that C is shattered by $mathcalH$ which means:

   $\text{VCdim}(\mathcal{H}_{=k}) \geq \min\{k, |\mathcal{X}| - k\}$ so $\text{VCdim}(\mathcal{H}_{=k}) = \min\{k, |\mathcal{X}| - k\}$

2. $\mathcal{H}_{at-most-k} = \left\{ h \in \{0,1\}^{\mathcal{X}} : |\{x : h(x) = 1\}| \leq k \ or \ |\{x : h(x) = 0\}| \leq k \right\}$

   Let $C \subseteq \mathcal{X}$ such that $|C| = k+1 \Rightarrow \exists h \in \mathcal{H}$ s.t $h_C(x) = 1 \Rightarrow \text{VCdim}(\mathcal{H}_{at-most-k}) \leq k$ ①

   Let $C \subseteq \mathcal{X}$ such that $|C| = m \leq k, C = x_1, ..., x_2$ With the same inference as the previous question we have $\text{VCdim}(\mathcal{H}_{at-most-k}) \geq k$ ②

   From ① and ② we conclude that $\text{VCdim}(\mathcal{H}_{at-most-k}) = k$

Exercise 6.6:(VC-dimension of Boolean conjunctions)

Problem assumptions: $\mathcal{H}_{con}^d \quad , \quad x_1, ..., x_d \quad , \quad s \geq 2$

1. There are three choice for each variable. (x_i,$\bar{x}_i$,None of them) $\Rightarrow |\mathcal{H}_{con}^d| = 3^d$
2. conclude that: $\text{VCdim}\left(\mathcal{H}_{con}^d\right) \leq \left\lfloor \log\left(\left|\mathcal{H}_{con}^d\right|\right) \right\rfloor \leq 3\log(d))$
3. We have to show that $\mathcal{H}_{con}^d$ shatters the set of unit vectors $\{e_i, i \leq d\}$
4. show that $\text{VCdim}\left(\mathcal{H}_{con}^d\right) \leq d$
5. $\mathcal{H}_{con}^d$ be monotone boolean conjunctions over $\{0,1\}^d$

   ✔We examine answers 3, 4 and 5 together.

   If the variable appears with its negation, it is assigned to -1 and otherwise to 1. $x_1 \wedge \bar{x}_1 \mapsto -1$

   $e_i = [0, 0, ..., \underbrace{1}_{i^{\text{th}}}, ..., 0, 0]$ We make the matrix X from the e_i vectors.

$$X = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \cdots & 1 \end{bmatrix}_{(d+1)\times(d+1)} \qquad Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_{d+1} \end{bmatrix} = \begin{bmatrix} \pm 1 \\ \pm 1 \\ \pm 1 \\ \vdots \\ \pm 1 \end{bmatrix}$$

This matrix is invertible with Determinant $d+1$.Therefore,$Xw = Y \Rightarrow w = X^{-1}Y$ So $d+1$ points can be shattered as a result: VCdim $\geq d+1$ ①

According to Radon's theorem, the $d+2$ points cannot be shattered. as a result:VCdim $\leq d+1$ ②

From ① and ②, we can conclude that VCdim $= d+1$

Exercise 6.9:

Let $\mathcal{H}$ be the class of signed intervals. $\mathcal{H} = \{h_{a,b,c} : a \leq b, s \in \{-1, 1\}\}$ when

$$h_{a,b,c}(x) = \begin{cases} s & \text{if } x \in [a,b] \\ -s & \text{if } x \notin [a,b] \end{cases}$$

We claim that VCdim($\mathcal{H}$) = 3. At the beginning we have to show that VCdim($\mathcal{H}$) ≥ 3 for this purpose let $C = \{x_1, x_2, x_3, x_4\}$ such that $x_i < x_i + 1$. We can easily find out that the label $(-1, +1, -1, +1)$ is not obtained by hypothesis $\mathcal{H}$

Exercise 6.10:

1. For every algorithm, there exists a distribution $\mathcal{D}$, for which $\min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) = 0$, but;

$$\mathbb{E}[L_{\mathcal{D}}(A(S))] \geq \frac{k-1}{2k}$$

$$k = \frac{d}{m} \Rightarrow \mathbb{E}[L_{\mathcal{D}}(A(S))] \geq \frac{d-m}{2d}$$

2. If $\mathcal{H}$ is PAC learnable, then VCdim($\mathcal{H}$) $< \infty$ Suppose if VCdim($\mathcal{H}$) $= \infty$ then $\mathcal{H}$ is not PAC learnable.

$$L_{\mathcal{D}}(A(S)) \leq \min L_{\mathcal{D}}(h) + \epsilon$$

Fundamental Theorems Of Machine Learning: If $\mathcal{H}$ is Agnostic PAC learnable $\Rightarrow$ $\mathcal{H}$ is PAC learnable.

So, we can conclude that $\Rightarrow \mathbb{P}\left[L_{\mathcal{D}}(A(S)) \geq \epsilon\right] < \delta$

Exercise 6.11:

$d = \max_i \mathrm{VCdim}(\mathcal{H}_i) \geq 3$ and also $\mathcal{H} = \cup_{i=1}^{r} \mathcal{H}_i$

1. we have to show that:

$$\mathrm{VCdim}\left(\cup_{i=1}^{r} \mathcal{H}_i\right) \leq 4d\log(2d) + 2\log(r)$$

According to definition, we have:

$$\tau_{\mathcal{H}}(k) \leq \sum_{i=1}^{r} \tau_{\mathcal{H}_i}(k) \Rightarrow \tau_{\mathcal{H}}(k) \leq rm^d \Rightarrow k \leq d\log(m) + \log(r)$$

Now, let $a \leq 0$ and $b > 0$. Then, $x > 4a\log(2a) + 2b \Rightarrow x \leq a\log(x) + b$. Therfore,

$$\Rightarrow k \leq 4d\log(2d) + 2\log(r)$$

2. It is clear that $\text{VCdim}(\mathcal{H}_1) = \text{VCdim}(\mathcal{H}_2) = d$ also $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_\in$ and $k \geq 2d+2$ we are trying to show that $\tau_{\mathcal{H}}(k) < 2^k$.

$$\begin{aligned}
\tau_{\mathcal{H}}(k) \leq \quad & \tau_{\mathcal{H}_1}(k) + \tau_{\mathcal{H}_2}(k) \\
& \leq \textstyle\sum_{i=0}^{d} \binom{k}{i} + \sum_{i=0}^{d} \binom{k}{i} \\
& = \textstyle\sum_{i=0}^{d} \binom{k}{i} + \sum_{i=0}^{d} \binom{k}{k-i} \\
& = \textstyle\sum_{i=0}^{d} \binom{k}{i} + \sum_{i=k-d}^{d} \binom{k}{i} \\
& \leq \textstyle\sum_{i=0}^{d} \binom{k}{i} + \sum_{i=d+2}^{d} \binom{k}{i} \\
& < \textstyle\sum_{i=0}^{d} \binom{k}{i} + \sum_{i=d+1}^{d} \binom{k}{i} \quad = \sum_{i=0}^{k} \binom{k}{i} = 2^k
\end{aligned}$$

chapter 9

Exercise 9.1:

primary assumptions: $\ell|h(\mathbf{x}, y) = |h(\mathbf{x}) - y|$ and $|c| = \min_{a \geq 0} a$ s.t $c \leq a$ and $c \geq a$

Frist we define a vector called a as $a = (a_1, ..., a_m)$ also, According to the *Hint*, $\min_{\mathrm{w}} \sum_{i=1}^{m} |\langle \mathrm{w}, \mathrm{x}_i \rangle - y_i|$ is equal to output of ERM (minimum). Therfore,

$$a_i \geq \langle \mathrm{w}, \mathrm{x}_i \rangle - y_i \Rightarrow \langle \mathrm{w}, \mathrm{x}_i \rangle - a_i \leq y_i$$
$$a_i \geq - \langle \mathrm{w}, \mathrm{x}_i \rangle + y_i \Rightarrow - \langle \mathrm{w}, \mathrm{x}_i \rangle - a_i \leq -y_i$$

$$\left.\begin{array}{cc}
A = [X - I_m; -X - I_m] & A \in \mathbb{R}^{2m \times (m+d)} \\
\mathbf{v} = (w_1, ..., w_d, s_1, ..., s_d) & \mathbf{v} \in \mathbb{R}^{d+m} \\
b = (y_1, ..., y_m, -y_1, ..., -y_m)^T & b \in \mathbb{R}^{2m} \\
\mathbf{c} = (\underbrace{0, \ldots, 0}_{d}, \underbrace{1, \ldots, 1}_{m}) & \mathbf{c} \in \mathbb{R}^{d+m}
\end{array}\right\} \Rightarrow \min \mathbf{c}^T \mathbf{v} \text{ s.t } A\mathbf{v} \leq b$$

$$\mathbf{c}^T \mathbf{v} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ \vdots \\ 1 \end{bmatrix} . \begin{bmatrix} w_1, & \ldots & , w_d, & a_1, & \ldots & , a_m \end{bmatrix} = (a_1 + \cdots + a_m) = \sum_{i=1}^{m} a_i$$

Exercise 9.3:

Theorem: $\frac{\langle \mathbf{w}^*, \mathbf{w}^{(T+1)} \rangle}{\|\mathbf{w}^*\| . \|\mathbf{w}^{(T+1)}\|} \geq \frac{\sqrt{T}}{RB}$ and also $T \leq (RB)^2$

$$R = \max \|\mathbf{x}_i\| \leq 1, \|\mathbf{w}^*\| = m \text{ for all } i \leq m \qquad y_i \langle \mathbf{w}^*, \mathbf{x}_i \rangle \geq 1$$
$$\Rightarrow B = \min \{\|\mathbf{w}\| : \forall i \in [m], y_i \langle \mathbf{w}, \mathbf{x}_i \ \rangle \geq 1\} \leq \sqrt{m} \Rightarrow (BR)^2 \leq m$$

$\forall i \in [d] : \text{Sign}(0) = -1$, $\text{Sign}(\langle \mathbf{w}, \mathbf{x} \rangle) = y_i$

$$y = \begin{bmatrix} \pm 1 \\ \vdots \\ \pm 1 \end{bmatrix}, y = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{bmatrix}, y = \sum_{j<i} e_j, \left\langle \mathbf{w}^{(i)}, \mathbf{x}_i \right\rangle \Rightarrow \text{So it shows the wrong label for everycase.}$$

We engage a vector which is called $\mathbf{w}^*$ equal to $\begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & 0 & \dots & 1 \end{bmatrix}$ with this problem to meet the requirements of the problem. Since it predicts labels correctly, so: $\mathbf{x}\mathbf{w} = y \Rightarrow \mathbf{w} = \mathbf{x}^{-1}y$. Therfore,

$$\mathbf{x}\mathbf{w} = y$$

Exercise 9.4:

Consider all positive examples of the form $(\alpha, \beta, 1)$; $\alpha^2 + \beta^2 + 1 \leq R^2$. Furthermore, $y \langle \mathbf{w}^*, \mathbf{x} \rangle \geq 1$ (linearly separable) We show a sequence of R^2 examples on which the Perceptron makes R^2 mistakes.

$$(\alpha_1, 0, 1); \alpha_1 = \sqrt{R^2 - 1}$$

Now, on round t^th let the new example be such that the following conditions hold:

$$\begin{cases} (a) & \alpha^2 + \beta^2 + 1 = R^2 \\ (b) & \langle \mathbf{w}_t, (\alpha, \beta, 1) \rangle = 0 \end{cases}$$

We show that if $t \leq R^2$ both conditions will be satisfied:

$$\begin{aligned} \mathbf{w}^{(t-1)} &= (a, b, t-1) \\ \|\mathbf{w}_{t-1}\| = (t-1)R^2 \Rightarrow a^2 + b^2 + (t-1)^2 &= (t-1)R^2 \\ (a, 0, t-1); a &= \sqrt{(t-1)R^2 - (t-1)^2} \\ \text{Then for every } B \\ \langle (a, 0, t-1), (\alpha, \beta, 1) \rangle &= 0 \\ \alpha + 1 \leq R^2 \Rightarrow \beta &= \sqrt{R^2 - \alpha^2 - 1} \\ \alpha^2 + 1 = \frac{(t-1)^2}{\alpha^2} + 1 &= \frac{(t-1)^2}{(t-1)R^2 - (t-1)^2} + 1 = \\ \frac{(t-1)R^2}{(t-1)R^2 - (t-1)^2} &= R^2 . \frac{1}{R^2 - (t-1)} \leq R^2 \end{aligned}$$

where the last inequality assumes $R^2 \geq t$