Homework #3

Due: 2025-12-24 23:59 | 6 Problems, 100 Pts

Name: 赵凌哲, ID: 2300015881

Problem 1 (15'). Find out and prove the VC-dimension of the hypothesis class $\mathcal{H}_n$ on instance space $\{0,1\}^n$ ($n \geq 1$) where

$$\mathcal{H}_n = \{\{\boldsymbol{x} \in \{0,1\}^n \mid f_S(\boldsymbol{x}) = -1\} \mid S \subseteq \{1,2,\cdots,n\}\}.$$

Here, $f_S(\boldsymbol{x}) : \{0,1\}^n \to \{-1,+1\}$ is defined as

$$f_S(\boldsymbol{x}) := \begin{cases} -1, & S = \varnothing; \\ (-1)^{\prod_{j\in S} x_j}, & S \neq \varnothing. \end{cases}$$

Express the answer as a function of n. ◀

Answer. 首先证明 $VC(H_n) >= n$，我们构造如下的 n 个点：

$$x^{(i)} = (1,1,\ldots,1,\underbrace{0}_{i\text{位}},1,\ldots,1), \qquad \forall i$$

注意到 $f_S(\boldsymbol{x})$ 起到类似析取 $\boldsymbol{x}$ 的全部位的作用，因此对于上述 n 个点的任意子集:

$$B \subseteq \{x^{(1)},\ldots,x^{(n)}\}$$
$$\text{可构造：} S = \{i : \ x^{(i)} \notin B\}$$

对于任意 i 不难得知：

$$x^{(i)} \in B \Longrightarrow i \notin S \Longrightarrow f_S(x^{(i)}) = -1$$
$$x^{(i)} \notin B \Longrightarrow i \in S \Longrightarrow f_S(x^{(i)}) = 1$$

因此 n 个点的任意子集都可以被如此打散。故 $VC(H_n) >= n$。

再证明 $VC(H_n) <= n$，注意到 $|H_n| = 2^n$。而如果打散 $m > n$ 个点，起码需要 2^m 种 S，无法实现。因此有 $VC(H_n) <= n$。

综上有 $VC(H_n) = n$。 ◁

Problem 2 (14'). The shatter function $\pi_{\mathcal{H}}(n)$ is the maximum number of subsets of any set A of size n that can be expressed as $A \cap h$ for $h \in \mathcal{H}$. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two hypothesis classes and $\mathcal{H} = \{h_1 \cap h_2 \mid h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\}$. Recall that we have proved $\pi_{\mathcal{H}}(n) \leq \pi_{\mathcal{H}_1}(n)\pi_{\mathcal{H}_2}(n)$ in class.

(1) (6') Recall the Sauer's lemma we have learned in class. Sauer's lemma tells that for a hypothesis class $\mathcal{H}$ with VC-dimension d, $\pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}$. Prove that $\sum_{i=0}^{d} \binom{m}{i} \leq \left(\frac{em}{d}\right)^d$ when $m \geq d$.

(2) (8') For a hypothesis class $\mathcal{H}$ with VC-dimension d, define the hypothesis class $\mathcal{H}^k$ $(k \geq 2)$ as

$$\mathcal{H}^k = \left\{ \bigcap_{i=1}^{k} h_i \;\middle|\; h_i \in \mathcal{H} \right\}.$$

Prove that, the VC dimension of $\mathcal{H}^k$ is no more than $7dk \ln k$. You may use the assertions above. $(\ln 2 \approx 0.693,\ \mathrm{e} \approx 2.718,\ \ln 7 \approx 1.946,\ \ln\ln 2 \approx -0.367)$

◀

Answer.

(1) 数学归纳法。若 $m = d$，则原式化为 $2^m \leq e^d$，显然成立。若 $m - 1 \geq d$，对 m 的大小进行归纳。容易验证 $m = 1, 2$ 的情况成立。假设对于 $1, 2, ..., m-1$ 皆有上述不等式成立。则：

$$\begin{aligned}\binom{m}{\leq d} &= \binom{m-1}{\leq d} + \binom{m-1}{\leq d-1} \\ &\leq \left(\frac{\mathrm{e}(m-1)}{d}\right)^d + \left(\frac{\mathrm{e}(m-1)}{d-1}\right)^{d-1}\end{aligned}$$

我们证明下式成立即可：

$$\begin{aligned}&\left(\frac{\mathrm{e}(m-1)}{d}\right)^d + \left(\frac{\mathrm{e}(m-1)}{d-1}\right)^{d-1} \leq \left(\frac{\mathrm{e}m}{d}\right)^d \\ \Longrightarrow &\left(\frac{d}{d-1}\right)^{d-1} \cdot \frac{d}{e}(m-1)^{d-1} \leq m^d - (m-1)^d\end{aligned}$$

又因为由二项式展开 $m^d - (m-1)^d \geq d \cdot (m-1)^{d-1}$，且 $\left(\frac{d}{d-1}\right)^{d-1} \leq e$。将两者分别代入，发现上不等式成立。得证。

(2) 课堂已经证明 $\pi_{H_1 \cap H_2}(m) \leq \pi_{H_1}(m)\pi_{H_2}(m)$，且有上一问结论 $\binom{m}{\leq d} \leq \left(\frac{\mathrm{e}m}{d}\right)^d$，直接将结论迭代 k 次可以得到：

$$\pi_{H_k}(m) \leq \big(\pi_H(m)\big)^k \leq \left(\frac{em}{d}\right)^{dk}$$

令 $D = \mathrm{VC}(H_k)$，根据定义我们有：

$$\begin{aligned}2^D &\leq \left(\frac{eD}{d}\right)^{dk} \\ D \ln 2 &\leq dk \ln\left(\frac{eD}{d}\right)\end{aligned}$$

下面证明若 $D \geq 7dk \ln k$ 会矛盾。令

$$\begin{aligned}&D = 7dk \ln k \\ \Longrightarrow \quad &\ln\left(\frac{eD}{d}\right) = 1 + \ln 7 + \ln k + \ln\ln k \\ \Longrightarrow &LHS = 7dk \ln k \cdot \ln 2, \quad RHS = dk\big(1 + \ln 7 + \ln k + \ln\ln k\big)\end{aligned}$$

希望证明 $\forall k \geq 2$，都有：

$$(7 \ln 2 \cdot \ln k) - (\ln k + \ln \ln k + \ln 7 + 1) > 0$$

计算知上式在 $k = 2$ 时取最小，此时用题给近似得到上式约等于 $0.088 > 0$。从而当 $D = 7dk \ln k$ 时，必有 $2^D \leq \left(\frac{eD}{d}\right)^{dk}$ 不成立。当 $D \geq 7dk \ln k$ 时更不成立。故与上述不等式产生矛盾。

综上有 $\mathrm{VC}(H_k) = D \leq 7dk \ln k$，得证。

◁

Problem 3 (16'). Consider the following randomized online learning algorithm for expert advice problem.

Algorithm 1: Randomized online learning algorithm for expert advice

Set a constant $\eta > 0$, the number of experts N

$\boldsymbol{L}_0 \leftarrow 0^N$ ▷ Cumulative loss vector.

for $t = 1, 2, \cdots, T$ **do**

  $W(t) = \sum_i \exp(-\eta \boldsymbol{L}_{t-1}(i))$ ▷ Normalization coefficient.

  Select the i-th expert with probability $\boldsymbol{p}_t(i) = \frac{\exp(-\eta \boldsymbol{L}_{t-1}(i))}{W(t)}$

  Observe the loss vector $\boldsymbol{l}_t \in [0,1]^N$ for each expert ▷ The loss is guaranteed in $[0, 1]$.

  Update the cumulative loss $\boldsymbol{L}_t \leftarrow \boldsymbol{L}_{t-1} + \boldsymbol{l}_t$

end

The expected loss is $\sum_{t=1}^{T} \boldsymbol{p}_t^\top \boldsymbol{l}_t$.

(1) (7') Prove that,

$$\frac{W(t+1)}{W(t)} = \boldsymbol{p}_t^\top \exp(-\eta \boldsymbol{l}_t).$$

(2) (9') Prove the following upper bound for expected loss:

$$\sum_{t=1}^{T} \boldsymbol{p}_t^\top \boldsymbol{l}_t - \boldsymbol{L}_T(i) \leq \frac{\ln N}{\eta} + T\eta$$

for any $i \in [N]$.

[Hint: Consider the potential function $\Phi_t = \frac{1}{\eta} \ln(W(t))$. You may find the following inequality useful: $\mathrm{e}^{-x} \leq 1 - x + x^2, x > 0$.]

◀

Answer.

(1) 由定义有:

$$
\begin{aligned}
W(t+1) &= \sum_{i=1}^{N} \exp(-\eta L_t(i)) \\
&= \sum_{i=1}^{N} \exp\bigl(-\eta(L_{t-1}(i) + l_t(i))\bigr) \\
&= \sum_{i=1}^{N} \exp(-\eta L_{t-1}(i)) \exp(-\eta l_t(i))
\end{aligned}
$$

根据 $p_t(i)$ 的定义有:

$$\exp(-\eta L_{t-1}(i)) = W(t)\, p_t(i)$$

将其代入则有:

$$W(t+1) = W(t) \sum_{i=1}^{N} p_t(i) \exp(-\eta l_t(i))$$

两边同除以 $W(t)$,即可得到欲证等式:

$$\frac{W(t+1)}{W(t)} = \sum_{i=1}^{N} p_t(i) \exp(-\eta l_t(i)) = p_t^\top \exp(-\eta l_t)$$

(2) 根据题目提示定义势函数:

$$\Phi_t = \frac{1}{\eta} \ln W(t)$$

由(1)可得

$$\Phi_{t+1} - \Phi_t = \frac{1}{\eta} \ln\left(\frac{W(t+1)}{W(t)}\right) = \frac{1}{\eta} \ln\bigl(p_t^\top \exp(-\eta l_t)\bigr)$$

用提示的 $e^{-x} \le 1 - x + x^2$ 来放缩上式中指数项,注意 $l_t(i) \in [0,1]$,有:

$$
\begin{aligned}
\exp(-\eta l_t(i)) &\le 1 - \eta l_t(i) + \eta^2 l_t(i)^2 \\
&\le 1 - \eta l_t(i) + \eta^2 l_t(i) \qquad (l_t(i)^2 \le l_t(i))
\end{aligned}
$$

左乘 $p_t^\top$ 后有:

$$
\begin{aligned}
p_t^\top \exp(-\eta l_t) &\le 1 - \eta p_t^\top l_t + \eta^2 p_t^\top l_t \\
&\le 1 - \eta p_t^\top l_t + \eta^2 \qquad (p_t^\top l_t \le 1)
\end{aligned}
$$

利用 $\ln(1+u) \le u$ 放缩得到:

$$\ln\bigl(p_t^\top \exp(-\eta l_t)\bigr) \le -\eta p_t^\top l_t + \eta^2$$

再代入势函数的式子:

$$\Phi_{t+1} - \Phi_t \le -p_t^\top l_t + \eta$$

对 $t = 1, \dots, T$ 求和，可得：

$$\Phi_{T+1} - \Phi_1 \le -\sum_{t=1}^{T} p_t^\top l_t + T\eta$$

移项得到

$$\sum_{t=1}^{T} p_t^\top l_t + \Phi_{T+1} \le \Phi_1 + T\eta$$

只需估算两个势函数的取值。

由 $L_0(i) = 0$ 知道：

$$W(1) = \sum_{i=1}^{N} e^0 = N, \quad \Phi_1 = \frac{\ln N}{\eta}$$

对任意专家 $\forall i \in [N]$：

$$W(T+1) = \sum_{j=1}^{N} e^{-\eta L_T(j)} \ge e^{-\eta L_T(i)}$$

$$\Longrightarrow \Phi_{T+1} = \frac{1}{\eta} \ln W(T+1) \ge -L_T(i)$$

两者皆代入上述不等式即可得到目标不等式：

$$\sum_{t=1}^{T} p_t^\top l_t - L_T(i) \le \frac{\ln N}{\eta} + T\eta$$

$\triangleleft$

Problem 4 (15'). Consider the following boosting algorithm we learned in class.

Algorithm 2: Boosting algorithm

Input: Number of iterations M (where M is odd), a sample S of n labeled examples $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ with labels $y_1, \cdots, y_n$, a γ-weak ($\gamma > 0$) learner (i.e., an algorithm that given n labeled examples and a non-negative weight $\boldsymbol{w} \in \mathbb{R}^n$, gives an hypothesis with at least $\frac{1}{2} + \gamma$ accuracy on the weight $\boldsymbol{w}$).

$\boldsymbol{w}_1 \leftarrow (1, 1, \cdots, 1)$ ▷ Initialize each example $\boldsymbol{x}_i$ to have a weight $\boldsymbol{w}_1(i) = 1$.

for $t = 1, 2, \cdots, M$ **do**
- Call the γ-weak learner on the sample S with weight $\boldsymbol{w}_t$ to get the hypothesis h_t.
- **for** $i = 1, 2, \cdots, n$ **do**
  - **if** $\boldsymbol{h}_t(x_i) \ne y_i$ **then**
    - $\boldsymbol{w}_{t+1}(i) \leftarrow \boldsymbol{w}_t(i) \cdot \frac{\frac{1}{2}+\gamma}{\frac{1}{2}-\gamma}$
  - **else**
    - $\boldsymbol{w}_{t+1}(i) = \boldsymbol{w}_t(i)$
  - **end**
- **end**

end

Output: The classifier $\mathrm{Maj}(h_1, \cdots, h_M)$.

Assume hypothesis h_t has error rate β_t on the weighted sample $(S, \boldsymbol{w}_t)$.

(1) (10') Suppose β_t is much less than $\frac{1}{2}-\gamma$. Then, after the booster multiples the weight of misclassified examples by $\alpha=\frac{\frac{1}{2}+\gamma}{\frac{1}{2}-\gamma}$, hypothesis h_t will still have error less than $\frac{1}{2}-\gamma$ under the new weights. This means that h_t could be given again to the booster (perhaps for several times in a row). Calculate, as a function of α and β_t, how many times in a row h_t could be given to the booster before its error rate rises to above $\frac{1}{2}-\gamma$.

(2) (5') Modify the boosting algorithm in the following way: During the iteration, multiply the weight of each example that was misclassified by h_t by $\alpha_t=\frac{1-\beta_t}{\beta_t}$, instead of $\alpha=\frac{\frac{1}{2}+\gamma}{\frac{1}{2}-\gamma}$. Prove that, $h_{t+1}\neq h_t$.

◀

Answer.

(1) 记 E 为 h_t 分错的样本集合，C 为分对的集合。在某一轮中：

$$\begin{aligned} W &= \sum_i w(i) \\ &= W_E + W_C \\ &= \beta W + (1-\beta)W \end{aligned}$$

一次权重更新后，只有错分样本的权重被乘以 α，从而：

$$W_E' = \alpha W_E = \alpha\beta W$$

$$W' = \alpha W_E + W_C = \alpha\beta W + (1-\beta)W = \big((1-\beta)+\alpha\beta\big)W$$

新权重下同一个 h_t 的错误率变为：

$$\beta' = \frac{W_E'}{W'} = \frac{\alpha\beta}{(1-\beta)+\alpha\beta}$$

连续重复使用 r 次后，错分样本的权重被乘以 α^r，其余不变，从而：

$$\beta_r = \frac{\alpha^r\beta}{(1-\beta)+\alpha^r\beta}$$

欲使：

$$\begin{gathered} \beta_r \le \frac{1}{2}-\gamma \\ \Longrightarrow \frac{\alpha^r\beta}{(1-\beta)+\alpha^r\beta} \le \frac{1}{2}-\gamma \end{gathered}$$

化简得到：

$$\begin{gathered} \alpha^r \le \frac{(\frac{1}{2}-\gamma)(1-\beta)}{(\frac{1}{2}+\gamma)\beta} \\ \Longrightarrow r \le \frac{\ln\left(\dfrac{(\frac{1}{2}-\gamma)(1-\beta_t)}{(\frac{1}{2}+\gamma)\beta_t}\right)}{\ln\alpha} \end{gathered}$$

考虑到 r 为整数，整理得到最多连续使用的次数为：

$$r_{\max} = \left\lfloor \frac{\ln\left(\dfrac{(1-\beta_t)}{\beta_t}\right)}{\ln\alpha} - 1 \right\rfloor$$

(2) 错分样本的权重被乘以 $\alpha_t = \frac{1-\beta_t}{\beta_t}$，和（1）中计算完全类似，同一个 h_t 在新权重下的错误率为：

$$\beta_t' = \frac{\alpha_t\beta_t}{(1-\beta_t)+\alpha_t\beta_t} = \frac{(1-\beta_t)}{(1-\beta_t)+(1-\beta_t)} = \frac{1}{2}$$

γ-weak learner 定义要求输出的假设在新权重下错误率不超过 $\frac{1}{2}-\gamma$。

根据上述计算，h_t 在新权重下的错误率等于 $\frac{1}{2}$，不满足弱学习器条件，因此：

$$h_{t+1} \neq h_t$$

◁

Problem 5 (20'). A bipartite graph is a graph whose vertices can be divided into two disjoint and independent sets U and V such that every edge connects a vertex in U to one in V. Find out and prove the threshold for $\mathcal{G}(n,p)$ to be bipartite.

[Hint: The definition of bipartite graph is equivalent to a graph that does not contain any odd-length cycles.] ◀

Answer. threshold 为 $\frac{1}{n}$，下面分别说明 $p < \frac{1}{n}$ 和 $p > \frac{1}{n}$ 的情况。

(1) 当 $p < \frac{1}{n}$：

设 X_ℓ 表示图中长度为 ℓ 的简单环的个数，其中 $\ell \geq 3$。则有：

$$\mathbb{E}[X_\ell] \leq \frac{n^\ell}{2\ell}p^\ell$$

上式即为 n choose ℓ 再除以环的 2ℓ 种等价表示。令 $X = \sum_{\ell\geq 3} X_\ell$ 表示图中所有环的总数，则

$$\mathbb{E}[X] \leq \sum_{\ell\geq 3}\frac{(np)^\ell}{2\ell}$$

当 $p < \frac{1}{n}$ 且 $n \to +\infty$ 时，有 $np \to 0$，上式严格小于一个首项、公比都趋于 0 的等比数列，故求和式仍为 0，从而：

$$\mathbb{E}[X] \to 0 \Longrightarrow \mathbb{P}(X \geq 1) \leq \mathbb{E}[X] \to 0 \quad \text{(Markov)}$$

因此 w.h.p. 图中不存在任何环，进而不存在奇环，从而 w.h.p. $\mathcal{G}(n,p)$ 是二分图。

(2) 当 $p > \frac{1}{n}$：

只需证明 w.h.p. 存在一个奇环。此处证明 w.h.p. 存在一个三角形。

设 X 表示图中三角形的个数。每个三角形由三个顶点组成，其三条边全部存在的概率为 p^3，因此

$$\mathbb{E}[X] = \binom{n}{3} p^3 \sim n^3 p^3$$

当 $p > \frac{1}{n}$ 且 $n \to +\infty$ 时，有 $np \to \infty$，因此 $\mathbb{E}[X] \to \infty$

下面估计方差。令

$$X = \sum_{\triangle} I_{\triangle}$$

其中 $I_{\triangle}$ 表示某个特定三角形是否存在，从而有：

$$\mathrm{Var}(X) = \sum_{\triangle} \mathrm{Var}(I_{\triangle}) + \sum_{\triangle \neq \triangle'} \mathrm{Cov}(I_{\triangle}, I_{\triangle'})$$

将方差拆为两项，放缩掉减去期望平方，则第一项满足：

$$\sum_{\triangle} \mathrm{Var}(I_{\triangle}) \leq \mathbb{E}[X]$$

第二项中，只有两个三角形共享一条边时协方差才非零。相当于 n choose 4 的数量级，每个此图形有 $6 - 1 = 5$ 条边。因此有：

$$\sum_{\triangle \neq \triangle'} \mathrm{Cov}(I_{\triangle}, I_{\triangle'}) \sim n^4 p^5$$

因此

$$\frac{\mathrm{Var}(X)}{\mathbb{E}[X]^2} \sim \frac{n^3 p^3 + n^4 p^5}{n^6 p^6} \to 0, \qquad \text{when } np \to \infty$$

由 Chebyshev 不等式，

$$\mathbb{P}(X = 0) \leq \mathbb{P}(|X - \mathbb{E}[X]| \geq \mathbb{E}[X]) \leq \frac{\mathrm{Var}(X)}{\mathbb{E}[X]^2} \to 0$$

因此 w.h.p. 存在三角形，从而存在奇环，$\mathcal{G}(n,p)$ 不是二分图。

综上知道 $\mathcal{G}(n,p)$ 是二分图的阈值为 $\frac{1}{n}$。

◁

Problem 6 (20'). A vertex is called an isolated vertex if it does not have any edges. Prove that, the threshold for $\mathcal{G}(n,p)$ of the existence of isolated vertex is $p = \frac{\ln n}{n}$. ◀

Answer. 下面分别说明 $p > \frac{\ln n}{n}$ 和 $p < \frac{\ln n}{n}$ 的情况：

(1) 当 $p > \frac{\ln n}{n}$，

设随机变量 X 表示图中孤立点的总数。

对任意顶点 v，它是孤立点当且仅当与其余 $n-1$ 个顶点之间的边全部不存在：

$$\mathbb{P}(v \text{ 孤立}) = (1-p)^{n-1}$$

$$\mathbb{E}[X] = n(1-p)^{n-1}$$

由 $(1-p)^{n-1} \le e^{-p(n-1)}$，得到：

$$\mathbb{E}[X] \le n\exp(-p(n-1)) \sim n\exp\big(-(\ln n + \omega(1))\big) = \exp(-\omega(1)) \to 0$$

由 Markov 不等式有：

$$\mathbb{P}(X \ge 1) \le \mathbb{E}[X] \to 0$$

因此 w.h.p. $\mathcal{G}(n,p)$ 中不存在孤立点。

(2) 当 $p > \frac{\ln n}{n}$，

在该条件下，

$$\mathbb{E}[X] \sim n\exp\big(-(\ln n - \omega(1))\big) = \exp(\omega(1)) \to \infty$$

下面计算方差来构造 Chebyshev 不等式证明。

对两个不同顶点 $u \neq v$，它们同时是孤立点需要总共 $(n-2)+(n-2)+1 = 2n-3$ 条边不存在：

$$\mathbb{P}(u,v \text{ 同时孤立}) = (1-p)^{2n-3}$$

$$\mathbb{E}[X(X-1)] = n(n-1)(1-p)^{2n-3}$$

同时我们有，

$$\mathbb{E}[X]^2 = n^2(1-p)^{2n-2}$$

因此有：

$$\mathrm{Var}(X) = \mathbb{E}[X] + \mathbb{E}[X(X-1)] - \mathbb{E}[X]^2 = \mathbb{E}[X] + \mathbb{E}[X]^2\left(\frac{n(n-1)}{n^2}\cdot\frac{1}{1-p} - 1\right) = \mathbb{E}[X] + \mathbb{E}[X]^2 \cdot o(1)$$

从而有：

$$\frac{\mathrm{Var}(X)}{\mathbb{E}[X]^2} = \frac{1}{\mathbb{E}[X]} + o(1) \to 0$$

由 Chebyshev 不等式，

$$\mathbb{P}(X=0) \le \mathbb{P}(|X - \mathbb{E}[X]| \ge \mathbb{E}[X]) \le \frac{\mathrm{Var}(X)}{\mathbb{E}[X]^2} \to 0$$

因此 $\mathcal{G}(n,p)$ 中 w.h.p. 存在孤立点。

综上我们证明了不存在孤立点的阈值为 $p = \frac{\ln n}{n}$

$\lhd$