۱) به صورت کلی ثابت می‌کنیم VC dimension برای Perceptron در فضای d بعدی، d+1 است. در نتیجه کوچکترین نقطه شکست برابر d+2 خواهد بود.

(I) $VCdim(Perception) \geq d+1$

$$f(x) = \begin{cases} 1 & w^T x + b \geq 0 \\ -1 & \text{else} \end{cases} \Rightarrow x^{(0)} = (0,\dots,0)^T,\ x^{(i)} = (0,\dots,1,0,\dots,0)^T,\ X = \{x^{(i)} \mid d \geq i \geq 0\}$$

اندیس iام

$$,\ y = (y_0,\dots,y_d)^T \in \{1,-1\}^{d+1}$$

$$\Rightarrow b = \frac{1}{2} y_0,\ w = (w_1,\dots,w_d),\ w_i = y_i\ (1 \leq i \leq d) \Rightarrow$$

$$w^T x + b = \frac{1}{2} y_0 + \sum_{j=1}^{d} y_j x_j \overset{x = x^{(i)}}{\Longrightarrow} w^T x^{(i)} + b = \begin{cases} \frac{1}{2} y_0 & i = 0 \\ y_i & i \neq 0 \end{cases}$$

d+1 نقطه را پوشش می‌دهد

(II) $VCdim(Perception) < d+2 \qquad w^T x + b = W^T X$

$$\Rightarrow \exists\, x^{(1)},\dots,x^{(d+2)} \in \mathbb{R}^d \Rightarrow X^T = (x^T, 1),\ W^T = (w^T, b),\ x^{(i)} \leftrightarrow X^{(i)}$$

که هر مقدار دهی آن‌ها پوشش داده می‌شود

$$X \in \mathbb{R}^{d+1}$$

$$\Rightarrow X^{(1)},\dots,X^{(d+2)} \in \mathbb{R}^{d+1} \Rightarrow \exists i : X^{(i)} = \sum_{j \neq i} a_j X^{(j)},\quad a_j \neq 0$$ حداقل یک

$$\Rightarrow S = \{j \mid j \neq i, a_j \neq 0\} \Rightarrow \forall j \in S : sign(a_j) = y^{(j)},\quad y^{(i)} = -1$$

label

$$\Rightarrow h(X) = \begin{cases} 1 & f(X) > 0 \\ -1 & \text{else} \end{cases}$$ ، طبق فرض خلف یک W وجود دارد که d+2 نقطه را پوشش می‌دهد

$$\forall j \in S : a_j \cdot W^T X^{(j)} > 0,\quad W^T X^{(i)} \leq 0 \quad (*)$$

$$,\ W^T X^{(i)} = W^T\Big(\sum_{j \neq i} a_j X^{(j)}\Big) = \sum_{j \in S} a_j \cdot W^T X^{(j)} > 0 \quad (**)$$

پس W وجود ندارد که $d+۲$ نقطه را در بر بگیرد

I = II $\Rightarrow$ $VC\ dim(Perceptron) = d+1 \Rightarrow lowest\ break\ point = d+۲$

جواب گزینه $\underline{5}$

۲) الف) ادعا می‌کنیم $VCdim(H) = n$

(I) $VCdim(H) \geq n$:

مجموعه $\{e_1, e_2, \dots, e_n\}$ را در نظر بگیرید. $e_i = \langle 00\cdots1\cdots0 \rangle \longrightarrow$ اندیس i ام

به ازای هر مقداردهی b به این n بردار، می‌خواهیم یک h معرفی کنیم که مناسب باشد.

$$\Rightarrow \quad b_1 \leftrightarrow e_1 \,,\; b_2 \leftrightarrow e_2 \,,\; \dots \,,\; b_n \leftrightarrow e_n \,,$$

$$\Rightarrow S = \{ i \mid b_i = 1 \} \Rightarrow h_S(e_i) = \begin{cases} 1 & i \in S \\ 0 & i \notin S \end{cases} = \begin{cases} 1 & b_i = 1 \\ 0 & b_i = 0 \end{cases} \Rightarrow VCdim(H) \geq n \checkmark$$

(II) $VCdim(H) < n+1$:

به ازای هر $n+1$ بردار، یک بردار (حداقل) وجود خواهد داشت که ترکیب خطی مابقی بردارهاست:

$$a_j = e_1 \oplus e_2 \oplus \dots \oplus e_k \Rightarrow b_1 = b_2 = \dots = b_k = 0 ,\; b_j = 1 \Rightarrow VCdim(H) < n+1$$

غیر ممکن است.

$$\text{(I)},\ \text{(II)} \Rightarrow VCdim(H) = n$$

Scanned with CamScanner

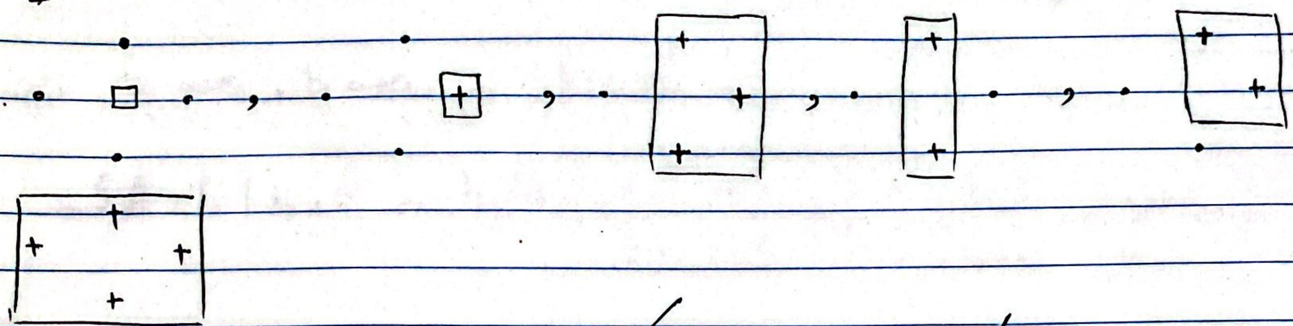
ب) ادعا می‌کنیم $VCdim(H) = 2d$ ،

(I) $VCdim(H) \geq 2d$

$e_i^+ = [0\,0 \ldots 1 \ldots 0]$ ، $e_i^- = [0\,0 \ldots -1 \ldots 0] \Rightarrow E^+ = \{e_i^+\}, E^- = \{e_i^-\} \Rightarrow S = E^+ \cup E^-$

اندیس iام ، اندیس iام

از $d=2$ آغاز می‌کنیم:



اکنون به ازای هر بعد جدید، d' ، $e_{d'}^+$ و $e_{d'}^-$ را اضافه می‌کنیم.

توجه کنید که به همان نتایج فرض بدست آمده برای $d'-1$ ، اضلاع لازم را اضافه می‌کنیم تا d' را نیز در بر بگیرد.

از آنجا که $e_{d'}^+$ و $e_{d'}^-$ در صفحه حاوی $\{e_1^+, \ldots, e_{d'-1}^+, e_1^-, \ldots, e_{d'-1}^-\}$ نیستند، اضافه کردن اضلاع به مستطیل جهت در بر گرفتن نقاط جدید، همواره ممکن است. (بسته به مقداردهی های مختلف به $e_{d'}^+$ و $e_{d'}^-$ ، حالات مختلفی پدید می‌آید)

به ازای $d=2$ برقرار و اگر برای d برقرار باشد، برای $d+1$ هم برقرار است $\Rightarrow VCdim(H) \geq 2d$

(1) (2)

(II) $VCdim(H) < 2d+1$

به ازای هر $\underbrace{2d+1 \text{ نقطه}}_{X}$ در فضا، ماکسیمم و مینیمم در ابعاد مختلف را محاسبه می‌کنیم.

$\rightarrow MIN = \{m_1, m_2, \ldots, m_d\}$ ، $MAX = \{M_1, \ldots, M_d\}$ ، $m_i = \min_{x \in X} x_i$ ، $M_i = \max_{x \in X} x_i$

$\Rightarrow \exists x \,\forall k : m_k \leq x_k \leq M_k \Rightarrow$ اگر x_k برابر (-1) مقداردهی شود و مابقی x_iها برابر $(+1)$، آنگاه امکان پذیر نیست چون x_k داخل مستطیل خواهد بود.

(I), (II) $\Rightarrow VCdim(H) = 2d$

ج)

$H_1 = \{ sign(p) \mid p: \mathbb{R}^d \to \mathbb{R} \} \longrightarrow$ به ازای هر n نقطه در فضای $\mathbb{R}^d$، بدیهتاً وجود خواهد p ای که
n نقطه را پوشش دهد (تابعی را در نظر بگیرید که دقیقاً همان نقاط را به $\{+1, -1\}$ مپ می‌کند).

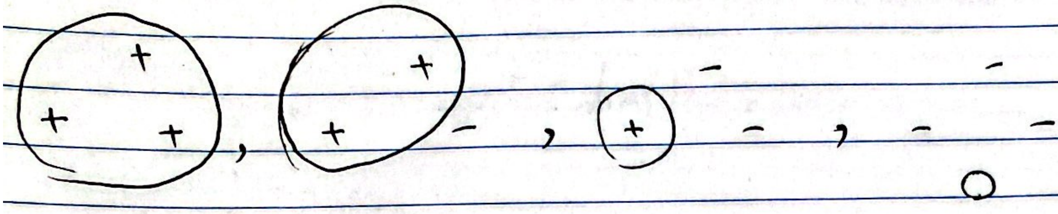
$\Rightarrow VCdim(H_1) = \infty$

و $H = \{ sign(f+g) \mid g: \mathbb{R}^d \to \mathbb{R}, f \in F \} \Rightarrow H_1 \subseteq H$
طبق تعریف

$\Rightarrow VCdim(H) = \infty$

د) ادعا می‌کنیم $VCdim(H) = 3$.

I) $VCdim(H) \geq 3$



پس به ازای ۳ نقطه بالا، فضای فرض هر مقداردهی را پوشش می‌دهد.

II) $VCdim(H) < 4$

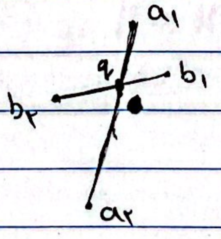
به ازای هر ۴ نقطه در فضا:

۱) همه آن‌ها روی یک خط باشند: + − + −

۲) ۳ تا از نقاط روی یک خط باشند: − / + − +

۳) هیچکدام روی یک خط نباشند و به صورت convex نباشند: + − + +

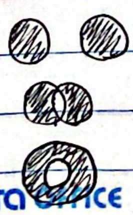
۴) هیچکدام روی یک خط نباشند و به صورت convex باشند:



ادعا می‌کنیم یکی از دو حالت $\underbrace{\{a_1:+, a_2:+, b_1:-, b_2:-\}}_{A}$ ، $\underbrace{\{a_1:-, a_2:-, b_1:+, b_2:+\}}_{B}$ ممکن نیست.

برهان خلف ⟵ اگر هر دوی اینها ممکن باشند: C_1 دایره مربوط به حالت ۱، C_2 دایره مربوط به حالت ۲ $\Rightarrow S = (C_1/C_2) \cup (C_2/C_1)$ را در نظر بگیرید.

⟸ S شامل ۴ ناحیه disjoint است. (وضعیت دو دایره نسبت به هم):



و در هیچکدام، ۴ ناحیه غیرمشترک وجود ندارد) ✗ ⟵ (۱)(۲)(۳)(۴) $\Rightarrow VCdim(H) < 4$

مسئله‌ی ۲.

پاسخ.

الف) فرض کنید که $\mathcal{E}(h_.) \leqslant \mathcal{E}(h^*) + \eta$ در آن صورت طبق نامساوی هافدینگ برای $\gamma = \sqrt{\frac{1}{2m}\log\frac{2k}{\delta}}$ با احتمال حداقل $1-\delta$ داریم که

$$\hat{\mathcal{E}}(h_.) \leqslant \mathcal{E}(h_.) + \gamma \leqslant \mathcal{E}(h^*) + \gamma + \eta \leqslant \mathcal{E}(\hat{h}) + \gamma + \eta$$
$$\leqslant \hat{\mathcal{E}}(\hat{h}) + 2\gamma + \eta$$

نامساوی های اول و آخر بر این اساس هستند که تمامی فرض ها در H باید در بازه خطای γ از خطای تعمیم اصلی خود باشند. نامساوی دوم بر اساس فرض سوال است. نامساوی سوم بر این فرض است که h^* کمینه کننده خطای تعمیم واقعی است و در نتیجه $\mathcal{E}(h^*) \leqslant \mathcal{E}(\hat{h})$.

با توجه به این موارد در نتیجه شرط برعکس یعنی

$$\hat{\mathcal{E}}(h_.) > \hat{\mathcal{E}}(\hat{h}) + 2\gamma + \eta$$

با احتمال حداکثر δ روی می‌دهد.

ب)

فرض کنید که $\mathcal{E}(h_.) > \mathcal{E}(h^*) + \eta$

در آن صورت طبق نامساوی هافدینگ برای $\gamma = \sqrt{\frac{1}{2m}\log\frac{2k}{\delta}}$ با احتمال حداقل $1-\delta$ داریم که

$$\hat{\mathcal{E}}(h_.) \geqslant \mathcal{E}(h_.) - \gamma \geqslant \mathcal{E}(h^*) - \gamma + \eta \geqslant \mathcal{E}(h^*) - 2\gamma + \eta$$
$$\geqslant \hat{\mathcal{E}}(\hat{h}) - 2\gamma + \eta$$

نامساوی های اول و آخر بر این اساس هستند که تمامی فرض ها در H باید در بازه خطای γ از خطای تعمیم اصلی خود باشند. نامساوی دوم بر اساس فرض سوال است. نامساوی سوم بر این فرض است که $\hat{h}$ کمینه کننده خطای empirical است.

با توجه به این موارد در نتیجه شرط برعکس یعنی

$$\hat{\mathcal{E}}(h_.) < \hat{\mathcal{E}}(\hat{h}) - 2\gamma + \eta$$

با احتمال حداکثر δ روی می‌دهد.

ج) فرض کنید که $h_. = h^*$. با استفاده از نامساوی هافدینگ که $\gamma = \sqrt{\frac{1}{2m}\log\frac{2k}{\delta}}$ با احتمال حداقل $1-\delta$ داریم که

$$\hat{\mathcal{E}}(h_.) \leqslant \mathcal{E}(h_.) + \gamma = \mathcal{E}(h^*) + \gamma$$
$$\leqslant \mathcal{E}(\hat{h}) + \gamma \leqslant \hat{\mathcal{E}}(\hat{h}) + 2\gamma$$

نامساوی های اول و آخر بر این اساس هستند که تمامی فرض ها در H باید در بازه خطای γ از خطای تعمیم اصلی خود باشند. تساوی دوم براساس فرض است. سومین نامساوی بر این اساس است که h^* کمینه کننده خطای تعمیم واقعی است. حال توجه کنید که برای η و δ ثابت، اگر $m \to \infty$ داریم $\gamma = \sqrt{\frac{1}{2m}\log\frac{2k}{\delta}} \to 0$

این بدین معنی است که برای m به اندازه کافی بزرگ $4\gamma < \eta$ و معادلا $2\gamma < \eta - 2\gamma$. در نتیجه با احتمال حداقل $1 - \delta$ اگر m به اندازه کافی بزرگ باشد $\hat{\mathcal{E}}(h_.) \leqslant \hat{\mathcal{E}}(\hat{h}) + \eta - 2\gamma$ در نتیجه الگوریتم جواب YES بر می‌گرداند.

6. **[10 points] Learning theory: Relaxed generalization bounds**

Let $Z_1, Z_2, \ldots, Z_m$ be independent and identically distributed random variables drawn from a Bernoulli(ϕ) distribution where $P(Z_i = 1) = \phi$ and $P(Z_i = 0) = 1 - \phi$. Let $\hat{\phi} = (1/m)\sum_{i=1}^m Z_i$, and let any $\gamma > 0$ be fixed. Hoeffding's inequality, as we saw in class, states

$$\mathbb{P}(|\phi - \hat{\phi}| > \gamma) \leq 2\exp(-2\gamma^2 m)$$

However, this relies on the assumption that the random variables $Z_1, \ldots, Z_m$ are all *jointly* independent. In this problem we will relax this assumption by only assuming *pairwise* independence among the Z_i. In this case we cannot apply Hoeffding's inequality, but the following inequality (Chebyshev's inequality) holds:

$$P(|\phi - \hat{\phi}| > \gamma) \leq \frac{\mathrm{Var}(Z_i)}{m\gamma^2}$$

where $\mathrm{Var}(Z_i)$ denotes the variance of the random variable Z_i and for $Z_i \sim \mathrm{Bernoulli}(\phi)$ we have $\mathrm{Var}(Z_i) = \phi(1-\phi)$.

Given our hypothesis set $\mathcal{H} = \{h_1, \ldots, h_k\}$ and m pairwise but not necessarily jointly independent data samples $(x, y) \sim \mathcal{D}$, we now derive guarantees on the generalization error of our best hypothesis

$$\hat{h} = \operatorname*{argmin}_{h \in \mathcal{H}} \hat{\varepsilon}(h)$$

where as usual we define $\hat{\varepsilon}(h) = \frac{1}{m}\sum_{i=1}^m 1\{h(x^{(i)}) \neq y^{(i)}\}$, where $(x^{(i)}, y^{(i)})$ are examples from the training set.

(a) [2 points] What is the maximum possible value of $\mathrm{Var}(Z_i) = \phi(1-\phi)$? From now on we will instead use this maximal value such that the bounds we derive hold for all possible ϕ.

Answer: We find the maximum value by using the first and second order conditions. Differentiating and setting to 0 gives $\phi = 1/2$. By finding the second derivative (-2), we confirm that this point is a maximum. Hence we substitute $\mathrm{Var}(Z_i)$ with $1/4$ for the remainder of the question.

(b) [4 points] Let $\gamma > 0$.

i. [2 points] Give a non-trivial (i.e. not the constant 1) upper bound on the probability that $|\hat{\varepsilon}(\hat{h}) - \varepsilon(\hat{h})| > \gamma$.

ii. [1 points] Fix $\delta \in (0,1)$. Using your upper bound, how large must the sample size m be before you can guarantee that

$$\mathbb{P}(|\hat{\varepsilon}(\hat{h}) - \varepsilon(\hat{h})| > \gamma) \le \delta,$$

that is, that the training error and generalization error are within γ of one another with probability at least $1-\delta$?

iii. [1 points] How does this sample size compare to what is achievable using Hoeffding's inequality?

Answer: We first use the Union bound to find

$$\begin{aligned} P(\exists h \in \mathcal{H}, |\varepsilon(h) - \hat{\varepsilon}(h)| > \gamma) &\le \sum_{i=1}^{k} P(|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) \\ &\le \sum_{i=1}^{k} \frac{1}{4m\gamma^2} \text{ (using Chebyshev's inequality)} \\ &= \frac{k}{4m\gamma^2}. \end{aligned}$$

(Note that applying Chebyshev's inequality to $\hat{h}$ does *not* work.) Setting this equal to δ and solving for m, we find the solution:

$$m = \frac{k}{4\delta\gamma^2}$$

Hence the number of training examples required to make this guarantee is linear in k instead of logarithmic as when we used Hoeffding's inequality.

(c) [4 points] Show that with probability at least $1-\delta$, the difference between the generalization error of $\hat{h}$ and the generalization error of the best hypothesis in $\mathcal{H}$ (i.e. the hypothesis $h^\star = \text{argmin}_{h\in\mathcal{H}}\, \varepsilon(h)$) is bounded by $\sqrt{k/(m\delta)}$.

Answer: First we solve for γ in the bound we found in (b):

$$\gamma = \sqrt{\frac{k}{4m\delta}}$$

Let $h^* = \arg\min_{h\in\mathcal{H}} \varepsilon(h)$. By uniform convergence and the definition of $\hat{h}$ (see Lecture Notes 4, page 7),

$$\varepsilon(\hat{h}) \le \varepsilon(h^*) + 2\gamma$$

Hence $|\varepsilon(\hat{h}) - \varepsilon(h^*)| \le 2\gamma = \sqrt{k/(m\delta)}$ as desired.