«به نام خدا»

محسن ابراهیمی _ ۹۵۱۰۵۳۰۲ تمرین سری پنجم

سوال T20

الف) (1). برای ضرب داخلی که در فضای F صورت می‌گیرد، خاصیت symmetric داریم پس:

$$\langle \varphi(x), \varphi(x') \rangle = \langle \varphi(x'), \varphi(x) \rangle$$

$$\Rightarrow K(x, x') = K(x', x)$$

(2). می‌دانیم جواب بهینه مسئله در فضای F برابر است با:

$$w^* = \sum_{i=1}^{m} \alpha_i \varphi(x_i) \qquad i = 1, \dots, m, \ \alpha_i \in \mathbb{R}$$

پس داریم:

$$\|w^*\|^2 \geq 0 \Rightarrow \langle w^*, w^* \rangle = \langle \sum_{i=1}^{m} \alpha_i \varphi(x_i), \sum_{j=1}^{m} \alpha_j \varphi(x_j) \rangle \geq 0$$

$$\Rightarrow \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j \langle \varphi(x_i), \varphi(x_j) \rangle \geq 0$$

$$\Rightarrow \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j K(x_i, x_j) \geq 0.$$

$\Rightarrow$ K is positive semi-definite function.

ب)

$$\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j K(x_i,x_j) \overset{*}{=} \sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j G_{ij}$$

$$G_{ij} = K(x_i, x_j) \quad *$$

$$\Rightarrow \sum_{i=1}^{m}\alpha_i \sum_{j=1}^{m}\alpha_j G_{ij} = \underline{\alpha^T G \alpha \geqslant 0 \ , \ \alpha \in \mathbb{R}^m / 0}$$

G is real matrix $\Rightarrow$ G is positive semi-definite matrix.

سوال T21

الف)

$$l \leqslant d \Rightarrow l = 0 \text{ یا } 1 \text{ یا } \dots \text{ یا } d$$

$$\Rightarrow l = 0 \rightarrow 1$$

$$\Rightarrow l = 1 \rightarrow K \qquad \Rightarrow |\chi_d| = 1 + K + \dots + K^d$$

تصاعد هندسی

$$\vdots$$

$$l = d \rightarrow K^d \qquad \Rightarrow |\chi_d| = \frac{K^{d+1} - 1}{K - 1}$$

همان‌طور که می‌بینیم با افزایش d، مقدار $|\chi_d|$ بصورت نمایی با K^{d+1} افزایش می‌یابد در نتیجه خواهیم داشت که:

$$|\chi_d| = \underbrace{\frac{1}{1 - 1/K}}_{\text{Constant}} K^d - \underbrace{\frac{1}{K-1}}_{\text{کم‌نسبت}} \Rightarrow \boxed{|\chi_d| = O(K^d)}$$

ادامه سوال T21

(ب)

$$w^u = \begin{cases} 1 & u = v \\ 0 & o.w \end{cases} \Rightarrow W = (0, 0, \dots, 1, \dots, 0, 0 \dots 0)$$

مؤلفه مربوط به ویروس v در feature set.

$$\Rightarrow \|w\|^2 = \langle w, w \rangle = \langle (0, \dots, 1, \dots, 0), (0, \dots, 0, 1, \dots, 0) \rangle$$

$$= 1 \Rightarrow \boxed{\|w\| = 1}$$

حال برای $\text{margin} = \gamma = \frac{1}{2}$ کافی است b را برابر با $-\frac{1}{2}$ قرار دهیم در این صورت

خواهیم داشت:
• اگر x دارای زیردنباله v باشد در این صورت: $\psi^v(x) = 1$

• اگر x' ویروس v را در زیررشته های خود نداشته باشد: $\psi^v(x') = 0$

در این صورت داریم:

$$x: \underbrace{\langle w, \psi(x) \rangle}_{1} - \frac{1}{2} = \frac{1}{2} \geq \frac{1}{2} \checkmark$$

$$x': \underbrace{\langle w, \psi(x') \rangle}_{0} - \frac{1}{2} = -\frac{1}{2} \leq -\frac{1}{2} \checkmark$$

پس $(w, -\frac{1}{2})$ تعریف شده margin خواسته شده را دارند.

ج

$\|\psi(x)\|^2 = \langle \psi(x), \psi(x) \rangle = \#$ of substrings of x.

کافی است زیردنباله های x را محاسبه کنیم. برای هر i در x داریم:

$i:$ $1 \quad 2 \quad 3 \; - - - - - \; l$: $i \rightarrow$ موقعیت در رشته x

موقعیت	# زیردنباله
1	l
2	$l-1$
$\vdots$	
l	0

$$\Rightarrow \; l + (l-1) + \dots + 0 = \boxed{\frac{l(l-1)}{2}} < l^2$$

$$\Rightarrow \|\psi(x)\|^2 \leqslant l^2 \leqslant d^2 \Rightarrow \|\psi(x)\| \leqslant d$$

$$\Rightarrow \boxed{\|\psi(x)\| = O(d)}$$

(د)

با توجه به تابع ψ، فایل‌هایی که شامل ویروس v باشند به نقطه صفر محور مربوط v در F نگاشت می‌شوند و فایل‌هایی که دارای v باشند به نقطه 1 این محور پس مسئله حداقل با (w, b) معرفی شده در قسمت (ب) Seperable است.

حال چون مسئله seprable-case است از Hard-SVM استفاده می‌کنیم تا بیشترین margin درست بیابد.

(ه)

$$K(x, x') = \langle \psi(x), \psi(x') \rangle$$

حال مقدار $\psi(x)$ مؤلفه‌های برابر است با اینکه آیا زیردنباله v در x قرار دارد یا نه، حال داریم:

$$\langle \psi(x), \psi(x') \rangle = \psi^{v_1}(x) \cdot \psi^{v_1}(x') + \cdots + \psi^{v_s}(x) \cdot \psi^{v_s}(x')$$

حال مقدار هرکدام از عبارت بالا وقتی 1 است که دنباله v_i در هردوی آن‌ها وجود داشته باشد حال جمع آن‌ها برابر خواهد بود با تعداد زیردنباله‌های مشترکی که هردو فایل x و x' دارند.

$$K(x, x') = \langle \psi(x), \psi(x') \rangle : \#\ \text{of common substrings}$$

(و) کافی است چک کنیم که آیا زیردنباله‌های فایل x، در زیردنباله‌های فایل x' وجود دارند یا نه و تعداد آن‌ها را بشماریم. تعداد زیردنباله‌های x از $O(d^2)$ می‌باشد حال کافی است هرکدام از آن‌ها را با $O(d^2)$ تا زیردنباله فایل x' از $O(1)$ مقایسه کنیم که هزینه این عمل از $O(d^2)$ است. درکل حداکثر d^2 بار این کار را (برای عضوهای مختلف x) انجام می‌دهیم پس پیچیدگی الگوریتم حداکثر از $O(d^4)$ است.

ز) بله، با توجه به (w, b) ای که در قسمت «ب» تعریف کردیم داریم که برای هر $(x, y) \sim D^*$ ، $\psi(x)$ آن را به محور مربوط v ، نگاشت می‌کند که در آن محور، x های دارای ویروس به 1 و x های که ویروس ندارند به 0 ، نگاشت داده می‌شود که با توجه به آن داده‌ها seperable هستند.

$$w^* = w_{(\text{ب})} = \begin{cases} w^u = 1 & : u = v \\ w^u = 0 & : o.w \end{cases} \quad , \quad b^* = -\frac{1}{2}$$

$$\Rightarrow \min y_i \left(\langle w^*, \psi(x) \rangle + b^* \right) = \boxed{\frac{1}{2} = \gamma}$$

$$\max_{x \in \mathbb{R}^d} \|\psi(x)\|^2 = \max_{x \in \mathbb{R}^d} \langle \psi(x), \psi(x) \rangle = d^2$$

«قسمت ج»

$$\Rightarrow \|\psi(x)\| \leq d \Rightarrow \boxed{\rho = d}$$

ج) با توجه به اینکه مسئله به صورت خطی در فضای F قابل جداسازی است، خطای الگوریتم Hard-SVM بر روی داده آموزش S برابر با صفر خواهد بود:

$$\Rightarrow L_S(h) = \frac{1}{m}\sum_{i=1}^{m} \ell^{0-1}((w,b),(x_i,y_i))$$

طبق قضیه 15.4 در کتاب (w^*,b^*) مربوطه باید فرم همگن داشته باشند، بدین منظور (w,b) که در قسمت (ز) بدست آمد را بردن در فضای $\mathbb{R}^{d+1}$ و سپس Normalize کردن برای اینکه $\|w\|=1$ شود به فرم همگن تبدیل می‌کنیم:

$$w^* = (0,0,\dots,\underset{\text{مربوط } v}{1},0,\dots,0,\frac{-1}{2}) \Rightarrow \|w^*\| = \frac{\sqrt{5}}{2}$$

$$\Rightarrow \tilde{w} = \frac{w^*}{\|w^*\|} \Rightarrow \tilde{w} = (0,0,\dots,0,\frac{2}{\sqrt{5}},0,\dots,0,\frac{-1}{\sqrt{5}})$$

$$\xrightarrow{15.4} L_D(\tilde{w}) \leq \sqrt{\frac{4\left(\frac{d}{1/\sqrt{5}}\right)^2}{m}} + \sqrt{\frac{2\log(2/\delta)}{m}}$$

$$\leq \frac{2d}{\sqrt{5m}} + \sqrt{\frac{2\log(2/\delta)}{m}} \quad , \quad (\gamma = \frac{1}{\sqrt{5}} \ , \ \rho = d)$$

ط) با توجه به قسمت قبل، از نامساوی بالا استفاده می‌کنیم:

با احتمال حداقل 0.99 داریم:

$$\frac{2\times 10^3}{\sqrt{5}\sqrt{m}} + \frac{\sqrt{2\times \ln\frac{2}{10^{-2}}}}{\sqrt{m}} \leq \frac{1}{100}$$

$$\Rightarrow \sqrt{m} \geq \left(\frac{2\times 10^3}{\sqrt{5}} + \sqrt{2\ln\frac{2}{10^{-2}}}\right)\times 10^2 \simeq \boxed{900}$$

$$\Rightarrow \boxed{m \geq 900^2 = 810000}$$

س) از حد بالای بدست آمده در PAC learning استفاده می‌کنیم. چون الگوریتم یادگیری که استفاده کردیم به صورت خطی داده‌ها را جداسازی کرده است پس خطای تجربی آن صفر است که حداقل مقدار ممکن است، پس الگوریتم را می‌توان خروجی ERM درنظر گرفت و bound های بدست آمده را اعمال کرد:

$$m \geq \frac{\log\frac{1}{\delta} + VCdim(\mathcal{H})}{\epsilon} \cdot c$$

$$VCdim(\mathcal{H}) = |\mathcal{X}_d| + 1 = O(K^d) + 1 \leq 256^{1000} + 1$$

$$\Rightarrow m \geq \frac{\log 10^2 + 256^{1000} + 1}{10^{-2}} \cdot c \Rightarrow \boxed{m \geq 10^2 (256^{1000} + 6)}$$

ک)

$$h_v(x) = \begin{cases} 1 & v \text{ is a substring of } x. \\ 0 & \text{otherwise} \end{cases}$$

فرض کنیم هر برای $D \sim (x,y)$ بخواهیم ویروس را تشخیص دهیم، حال چون برای هر v ای در $\mathcal{H}$ h_v وجود دارد که به درستی ویروس در هر فایل تشخیص می‌دهد پس برای هر x از D نیز، فرضیه‌ای در $\mathcal{H}$ داریم که بدون خطا وجود ویروس را تشخیص می‌دهد پس داریم که $L_D(h_v)=0$ پس شرط Realizability برقرار است.

حال چون $\mathcal{H}$ محدود است Uniform Convergence داریم پس اگر الگوریتم یادگیری خود را روش ERM قرار دهیم، درنتیجه PAC learning خواهیم داشت و از آن

رابطه زیر بدست می‌آید:

$$|H| = |X_d| \overset{\text{با توجه قسمت}}{\underset{\text{الف}}{=}} O(K^d) \Rightarrow |H| \leqslant K^d$$

حال با احتمال حداقل $1-\delta$ داریم:

$$m_H(\epsilon,\delta) \leqslant \frac{\log\frac{|H|}{\delta}}{\epsilon} \leqslant \frac{\log\frac{K^d}{\delta}}{\epsilon}$$

$$\Rightarrow \boxed{m_H(\epsilon,\delta) \leqslant \log\frac{K^d}{\delta} / \epsilon}$$

اگر داشته باشیم: $K=256$, $d=1000$, $\epsilon,\delta = 10^{-2}, 10^{2}$:

$$m \geqslant m_H(\epsilon,\delta) \Rightarrow m \geqslant \frac{\log\frac{256^{1000}}{10^{-2}}}{10^{-2}} = \frac{\log 10^2 + \log 256^{1000}}{10^{-2}}$$

$$\Rightarrow m \geqslant \frac{2\log 10 + 1000\log 256}{10^{-2}} \simeq 56\times 10^4$$

نتیجه بدست آمده از حد پایین برای m، بند «ط» بسیار بزرگتر است اما نسبت به sample complexity که در قسمت «ی» بدست آوردیم کمبود چشمگیری داشته است.

Scanned by CamScanner

سوال T22

الف)

$$L_D(h) = \mathbb{E}_x \mathbb{E}_y \left[(y-h(x))^2 / x \right]$$

* از طرفی می‌دانیم:

$$\mathbb{E}_x \mathbb{E}_y \left[(y-h(x))^2 / x \right] = \sum_x \mathbb{E}_y \left[(y-h(x))^2 \mid x \right] \mathbb{P}_{x\sim D}(x) \quad (*)$$

برای کمینه کردن $L_D(h)$، کافی است مینیمم عبارت (*) را بدست آوریم. حال چون ما توزیع D مربوط به داده‌ها را می‌دانیم پس مقادیر $\mathbb{P}_{x\sim D}(x)$ برای x های مختلف از D را داریم پس درنتیجه برای مینیمم کردن $L_D(h)$، عبارت $\mathbb{E}_y\left[(y-h(x))^2/x\right]$ برای x ها مختلف مینیمم شود. (در واقع $L_D(h)$ به صورت ترکیب خطی از مقادیر $\mathbb{E}\left[(y-h(x))^2/x\right]$ می‌شود که برای کمینه کردن $L_D(h)$، باید هرکدام از این term ها کمینه شوند)

ب)

$$L = \mathbb{E}_y \left[(y-h(x))^2 / x \right] = \mathbb{E}_y \left[y^2 - 2y h(x) + h^2(x) / x \right]$$

$$= \mathbb{E}_y [y^2 / x] - 2h(x) \mathbb{E}_y [y / x] + h^2(x)$$

$$\Rightarrow \nabla_h L = 0 \Rightarrow \frac{\partial L}{\partial h} = 0 \Rightarrow -2 \mathbb{E}_y [y/x] + 2h(x) = 0$$

$$\Rightarrow h(x) = \mathbb{E}_y [y/x]$$

$$\Rightarrow \boxed{h_D(x) = \mathbb{E}_y [y/x]}$$

(ج)

$$h_D(x) = \underset{y}{\mathbb{E}}[y|x]$$

$$\Rightarrow \underset{y}{E}[(y-h_D(x))^2/x] = \underset{y}{\mathbb{E}}[(y-\underset{y}{\mathbb{E}}[y|x])^2|x]$$

$$= \underset{y}{\mathbb{E}}[y^2 - 2y\underset{y}{\mathbb{E}}[y|x] + (\underset{x}{\mathbb{E}}[y|x])^2 |x]$$

$$= \underset{y}{\mathbb{E}}[y^2/x] - 2\underset{y}{\mathbb{E}}[y|x]\underset{y}{\mathbb{E}}[y|x] + \underset{x}{\mathbb{E}}^2[y|x]$$

$$= \underset{y}{E}[y^2/x] - \underset{y}{\mathbb{E}}^2[y|x] = \text{Variance}(y|x) = \sigma^2_{y|x}$$

$$\Rightarrow L_D(h_D(x)) = \underset{x}{E}[\sigma^2_{y|x}] = \min_{h \in \mathcal{H}} L_D(h) = \epsilon_{bayes}$$

در بین تمام h ها ممکن

(د)

$$h_D(x) = \underset{y}{\mathbb{E}}[y|x] = y \times \underbrace{\underset{y\sim D}{\mathbb{P}}[y|x]}_{} = y = f(x)$$

deterministic رابطه $= 1$

$$\Rightarrow \underset{y}{\mathbb{E}}[(\overset{f(x)}{y} - h_D(x))^2/x] = \mathbb{E}[(y-y)^2/x] = 0 = \sigma^2_{y|x}$$

$$\Rightarrow L_D(h_D) = \underset{x}{\mathbb{E}}[\underbrace{\sigma^2_{y|x}}_{0}] = \emptyset$$

(ه)

$$h_D(x) = \underset{y}{\mathbb{E}}[y|x] = \sum_y y \cdot \underset{y\sim D}{\mathbb{P}}[y|x] = \sum_y y \cdot \frac{\mathbb{P}(x,y)}{\mathbb{P}(x)}$$

$$\xrightarrow[\text{مستقل اند}]{y,x \text{ از هم}} h_D(x) = \sum_y y \cdot \frac{\cancel{\mathbb{P}(x)}\mathbb{P}(y)}{\cancel{\mathbb{P}(x)}} = \sum_y y \cdot \underset{y\sim D}{\mathbb{P}}(y) = \boxed{\underset{y}{E}[y]}$$

$$\mathbb{E}_y\left[(y-h(x))^2 / x\right] = \mathbb{E}_y\left[(y-E[y]))^2 / x\right]$$

$$= \underbrace{\mathbb{E}_y[y^2/x]}_{\mathbb{E}[y^2]} - 2\mathbb{E}[y]\underbrace{\mathbb{E}[y/x]}_{\mathbb{E}[y]} + \underbrace{\mathbb{E}^2[y/x]}_{\mathbb{E}^2[y]}$$

$$= \mathbb{E}[y^2] - \mathbb{E}^2[y] = \sigma_y^2.$$

$$\Rightarrow L_D(h_D) = \mathbb{E}_x[\sigma_y^2] = \boxed{\sigma_y^2}$$

(و)

$$L_D(h) = \mathbb{E}_x \mathbb{E}_y \left[(h(x) - y)^2 \mid x \right], \quad h(x) = c$$

$$\Rightarrow L_D(h) = \mathbb{E}_x \mathbb{E}_y \left[(c - y)^2 \mid x \right] = \mathbb{E}_x \mathbb{E}_y \left[y^2 - 2cy + c^2 \mid x \right]$$

$$= \mathbb{E}_x \left[\mathbb{E}_y [y^2 \mid x] - 2c\, \mathbb{E}_y [y \mid x] + c^2 \right]$$

$$= \mathbb{E}_x \mathbb{E}_y [y^2 \mid x] - 2c\, \mathbb{E}_x \mathbb{E}_y [y \mid x] + c^2$$

$$\frac{dL_D}{dc} = 0 \Rightarrow -2\, \mathbb{E}_x \mathbb{E}_y [y \mid x] + 2c = 0$$

$$\Rightarrow \boxed{c = \mathbb{E}_x \mathbb{E}_y [y \mid x]} = \mathbb{E}_{x,y} [y]$$

$$\longrightarrow L_D(c) = \mathbb{E}_x \mathbb{E}_y \left[\left(\mathbb{E}_x \mathbb{E}_y [y \mid x] - y \right)^2 \mid x \right]$$

$$= \mathbb{E}_x \mathbb{E}_y [y^2 \mid x] - \left(\mathbb{E}_x \mathbb{E}_y [y \mid x] \right)^2$$

مقدار خطا به ازای c بهینه

(ز) رابطه صحیح به صورت مقابل است: $\mathbb{E}_x [\sigma^2_{y|x}] \leqslant \sigma^2_y$

وقتی حالت تساوی رخ می‌دهد که y و x از هم مستقل باشند اما وقتی رابطه استقلال ندارند بدین معناست که به ازای هر $x \sim D$، به یک توزیع وابسته به x، برای مقادیر y آن می‌رسیم یعنی $\mathbb{P}(Y|x)$ و سپس از روی این توزیع وابسته x، تخمین خود را که برابر با $\mathbb{E}_y [y \mid x]$ است، بدست می‌آوریم. اما وقتی استقلال برقرار باشد برای هر x، توزیع y ها وابسته به آن را نداریم و باید از روی خود توزیع y بهترین تخمین که برابر با $\mathbb{E}_y [y]$ است را در نظر بگیریم. که در نتیجه خطای بیشتری خواهیم داشت.

(e) $$\min_{h\in\mathcal{H}} L_D(h) \geq \epsilon_{bayes} = \mathbb{E}_x[\sigma^2_{y|x}]$$

بله، زیرا مقدار ϵ_{bayes} کمترین خطای است که به ازای همه‌ی h های ممکن بدست می‌آید

که شامل مجموعه فرضیه $\mathcal{H}$ نیز می‌شود، پس بهترین انتخاب ممکن از بین $\mathcal{H}$

حداقل کمترین خطای برابر با ϵ_{bayes} خواهد داشت.

(f) اگر شرط Realizability برروی $\mathcal{H}$ داشته باشیم بدین معناست که کمترین خطا براساس $\mathcal{H}$ برابر با صفر است:

$$\underbrace{\min_{h\in\mathcal{H}} L_D(h)}_{0} \geq \epsilon_{bayes} \geq 0$$

$$\Rightarrow \boxed{\epsilon_{bayes}=0}$$

(g) وقتی شرط Realizability برروی $\mathcal{H}$ برقرار است بدین معناست h ای در $\mathcal{H}$ وجود دارد که رابطه‌ی قطعی بین x و y در D برقرار می‌کند: $y=f(x)$

که با توجه به آن $\mathbb{E}_y[y|x]=y \Leftarrow \mathbb{P}(y|x)=1$

که در این صورت خواهیم داشت:

$$\mathbb{E}_x\Big[\mathbb{E}_Y\big(\underbrace{\overbrace{\mathbb{E}_y[y|x]}^{y}-y}_{0}\big)^2|X\big)\Big] = \mathbb{E}_X[0]=0$$

$$\Rightarrow \epsilon_{bayes}=0$$

در واقع با شرط Realizability حالت احتمالاتی مقادیر y برای x از بین می‌رود و برای هر x، یک y خواهیم داشت که با انتخاب آن، خطا روی کل D صفر می‌شود.

د) ابتدا: $(\epsilon_{app} - \epsilon_{bayes} \geq 0)$

ϵ_{bayes} کمترین خطای است که می‌توان از بین تمام hهای ممکن بدست آورد که مجموعه فرضیه $\mathcal{H}$ نیز عضو آن است. حال خطای بدست آمده برای بهترین h از $\mathcal{H}$، کوچکتر یا مساوی با ϵ_{bayes} خواهد بود:

$$\epsilon_{app} = \min_{h\in\mathcal{H}} L_D(h) \geq \epsilon_{bayes} = \min_{h:\,\text{تمام hهای ممکن}} L_D(h)$$

$$\Rightarrow \boxed{\epsilon_{app} - \epsilon_{bayes} \geq 0}$$

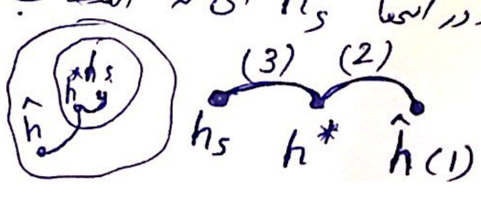
- ϵ_{bayes}: به این معناست که بهترین خطای که در تمام hها می‌توان برای مدل بدست آورد چقدر است. $(\hat{h})$

- $\epsilon_{app} - \epsilon_{bayes}$: اگر ما بهترین انتخاب h را از بین $\mathcal{H}$ انجام دهیم (h^*)، چقدر با بهترین h ممکن برای مدل فاصله داریم.

- ϵ_{est}: hای که براساس الگوریتم یادگیری و داده آموزشی که بدست می‌آوریم، مقدار خطای آن چقدر با مقدار خطای بهترین hای که می‌توانستیم از $\mathcal{H}$ انتخاب کنیم فاصله دارد.

~~همه~~ جمله اول (ϵ_{bayes}) در این جمع بیان می‌کند که حداقل چقدر خطا برای هر h ممکن دارم (شامل تمام hها از جمله h_s)، حال بهترین انتخاب h از $\mathcal{H}$ چقدر خطا نسبت به بهترین h ممکن دارد (جمله دوم) و در انتها h_s ای که انتخاب کرده‌ایم چقدر خطا نسبت به بهترین h در $\mathcal{H}$ دارد.

تمام hها



T23 سوال

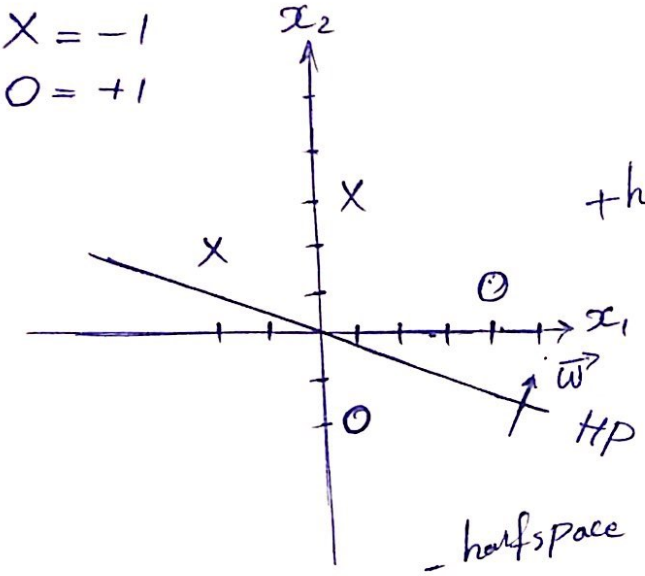
1

$$X = \mathbb{R}^2, \quad Y = \{-1, +1\}.$$

$$S = \{((-2,2),-1),\ ((1,3),-1),\ ((1,-2),+1),\ ((4,1),+1)\}.$$

$\times = -1$

$O = +1$



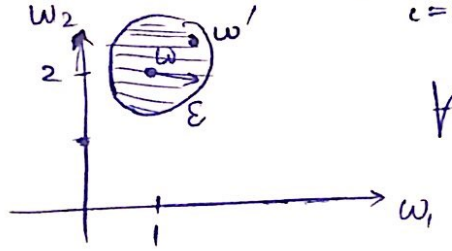
+halfspace

HP: $\langle (1,2), (x_1, x_2) \rangle = 0$

$h = \mathrm{sign}[x_1 + 2x_2]$

_halfspace $\Rightarrow$ $w = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $b = 0$

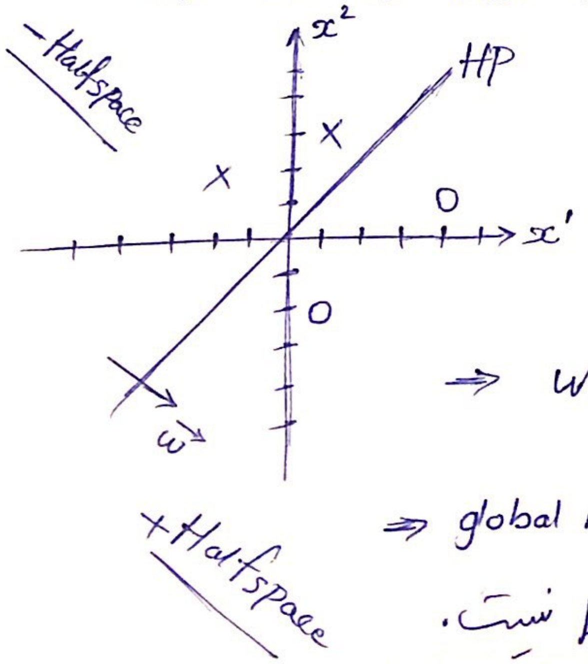
$$\Rightarrow L_S(w) = \frac{1}{4} \sum_{i=1}^{4} \ell^{0-1}(w, (x_i, y_i)) = \frac{1}{4}[+1 + 1 + 0 + 1] = \frac{3}{4}$$



$\forall w'$: $|w' - w| < \varepsilon$, for some $\epsilon > 0$:

$$L_S(w') = L_S(w) = \frac{3}{4}$$

$\Rightarrow$ w is local minimum of $L_S(w)$.



2

HP: $\langle (+1,-1), (x_1, x_2) \rangle = 0$

$h = \mathrm{sign}[x_1 - x_2]$

$\Rightarrow w^* = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \Rightarrow \boxed{L_S(w^*) = 0}$

$\Rightarrow$ global minimum چون خطای کمتر از ۰ نداریم پس w^* یک global minimum

است و w، $L_S(w)$ global min نیست.