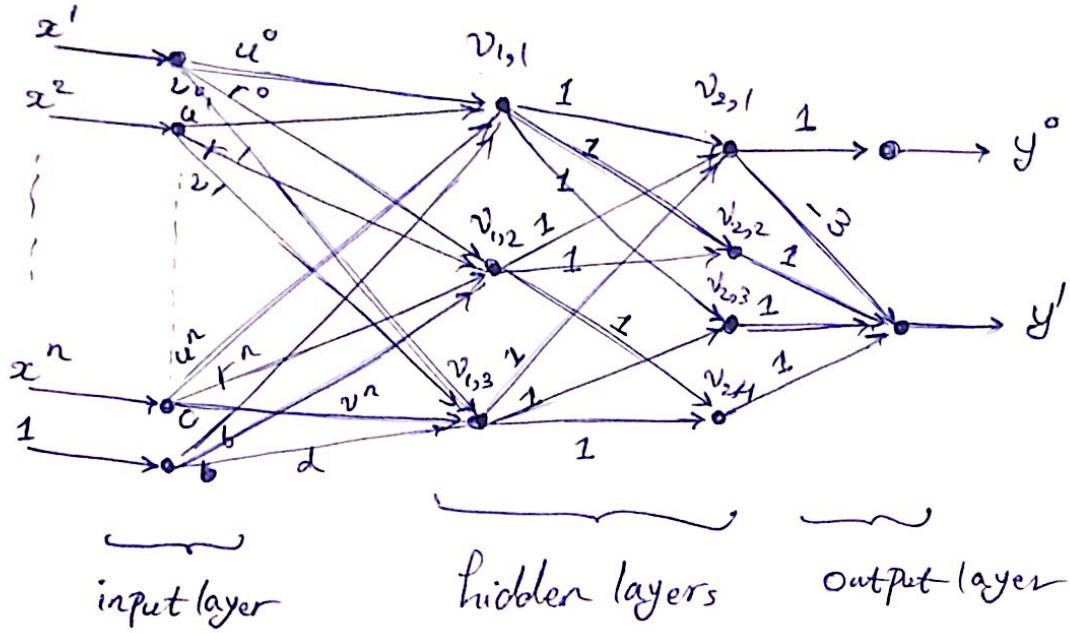
تمرین سری هفتم

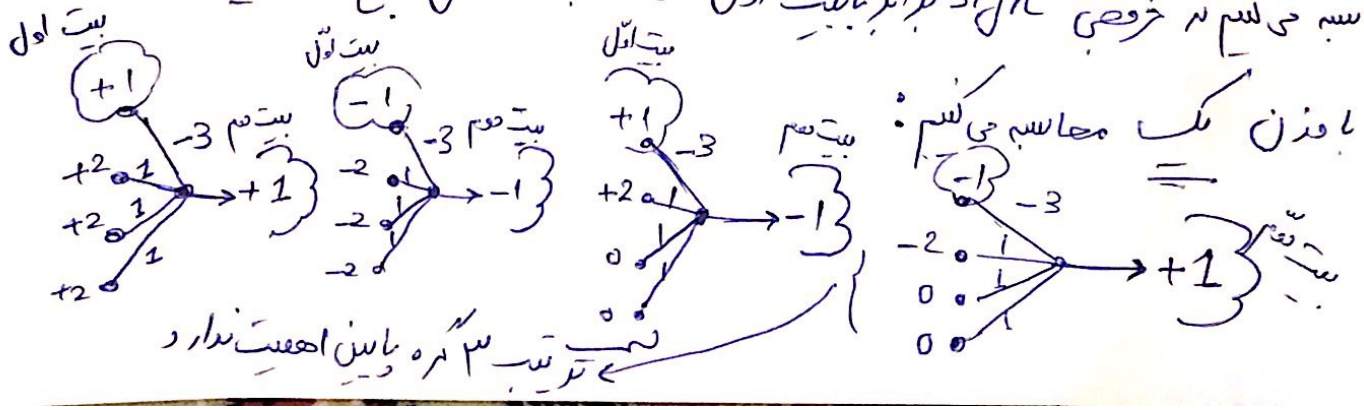
(بنام او)

محسن ابراهیمی - 95105302

سوال T8



در لایه اول hidden layer ، دو گره $v_{1,1}$ و $v_{1,2}$ و ورودی آن را برابر قرار می دهیم با همه ی ورودی ها x^1 تا x^n و 1 و وزن این یال ها را برابر قرار می دهیم با u^1 تا u^n و c . که با این حالت ترکیب خطی $(u^T x + c)$ محاسبه می شود در ورودی node ، $v_{1,1}$ و با استفاده از sign ، $\sigma = sign$ ، مقدار $sign(a^T x + c)$ در خروجی این گره قرار می گیرد. به طور مشابه این کار را برای گره های دیگر این لایه با وزن ها $(r_1 \dots r_n$ و $d)$ و $(v_1^1, v_2^1 \dots v_n^1$ و $d)$ انجام می دهیم. در لایه بعدی جمع 3 گره قبلی را با وزن های 1 محاسبه می کنیم که خروجی sign برابر با بیت اول خواهد بود. حال جمع دو تای ۳ گره لایه قبل را با وزن یک محاسبه می کنیم:



سوال T9

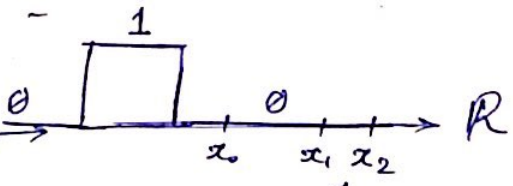
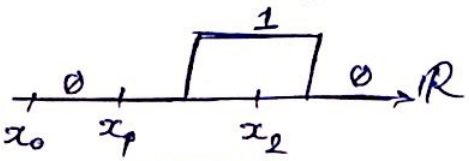
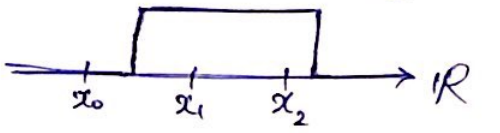
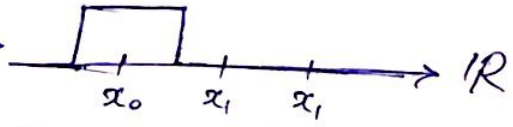
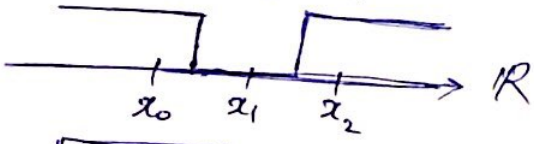
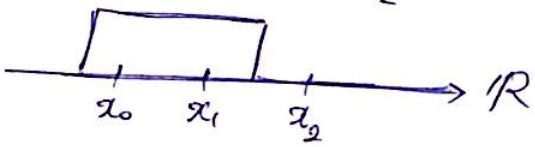
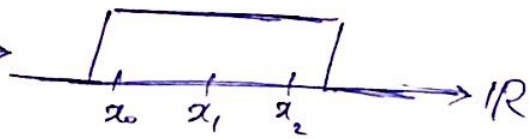
$$H = \{ h_{a,b,s} \;;\; a \le b,\ s \in \{-1,1\} \}.$$

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

حل: اگر تعداد نقاط خود را برابر با $d=3$ در نظر بگیریم داریم:

x_0, x_1, x_2

فرض: $x_0 < x_1 < x_2$

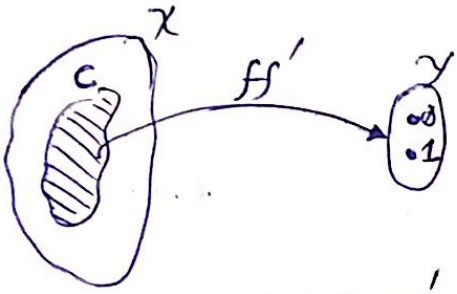
x_0	x_1	x_2	
0	0	0	
0	0	1	
0	1	0	
0	1	1	
1	0	0	
1	0	1	
1	1	0	
1	1	1	

طبق جدول بالا داریم که H بر $C = \{x_0, x_1, x_2\}$ غالب است پس $d \ge 3$. اما برای $C' = \{x_0, x_1, x_2, x_3\}$، $d = 4$ داریم که حالت متقارن دهی $(x_0, x_1, x_2, x_3) \sim (0,1,0,1)$ وجود ندارد. پس H اعضای $x_0 \le x_1 \le x_2 \le x_3$ فرض

$\overline{d=3}$ خواهد بود //

سوال T10

$$\mathcal{H}' \subseteq \mathcal{H} \longrightarrow VCdim(\mathcal{H}') \leq VCdim(\mathcal{H})$$



فرض کنید مجموعه C مجموعه‌ای باشد که $\mathcal{H}'$ بر آن غالب است و حال چون $\mathcal{H}' \subseteq \mathcal{H}$ است h های که در $\mathcal{H}'$ هستند تمام حالت‌های مقدار دهی C را پوشش می‌دهند در $\mathcal{H}$ نیز قرار دارند پس $\mathcal{H}$ نیز بر C غالب است و چون $\mathcal{H}$ اعضای بیشتری نسبت $\mathcal{H}'$ دارد ممکن C' با $|C'| > C$ وجود داشته باشد که $\mathcal{H}$ بر آن غالب باشد پس خواهیم داشت:

$$VCdim(\mathcal{H}') \leq VCdim(\mathcal{H})$$

سوال T11

$$X = \mathbb{R}^2, \quad \mathcal{H} = \{ h_{a_1,b_1,a_2,b_2} : a_1 < a_2 , b_1 < b_2 \}.$$

$$h_{a_1,a_2,b_1,b_2}(x) = \begin{cases} 1 & a_1 \leq x^1 \leq a_2 , b_1 \leq x^2 \leq b_2 \\ 0 & o.w \end{cases}$$

با توجه به اینکه $d = VCdim(\mathcal{H}) = 4$ که در قسمت اول بدست آمد و همچنین با زمانی که برای agnostic PAC learning داریم:

$$C_1 \frac{d + \log 1/\delta}{\epsilon^2} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{d + \log 1/\delta}{\epsilon^2}$$

$$\rightarrow \quad C_1 \frac{\log 1/\delta + 4}{\epsilon^2} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{4 + \log 1/\delta}{\epsilon^2} .$$

سوال T11 - (قسمت ج)

با استفاده از حد بالا ای که بدست آوردیم یعنی $m(\delta,\epsilon) \leq C_2 \frac{d+\log 1/\delta}{\epsilon^2}$ ، این اطلاعات را به ما می‌دهد که برای آنیله در حالت infinite class ، $\mathcal{H}$ ، agnostic PAC learnable باشد لازم نیست که تعداد نمونه های ما از کل جمعیت به سمت ∞ میل کند و می‌توان با متناهی حقیری کرد $m < \infty$ است، این خاصیت در مورد $\mathcal{H}$ داشت و حد پایین عبارت یعنی $m_{\mathcal{H}}(\epsilon,\delta) \geq C_1 \frac{d+\log \frac{1}{\delta}}{\epsilon^2}$ حاوی این اطلاعات است که با هر تعدادی نمونه از جمعیت نمی توان خاصیت مورد نظر را داشت و این تعداد نمونه ها باید از یک حدای که وابسته به ϵ و δ و d است بیشتر باشد.

سوال T12

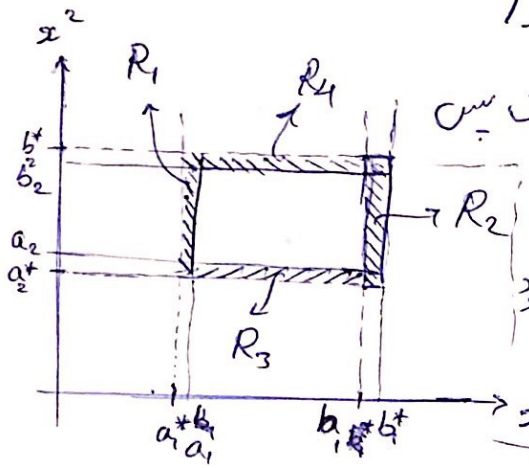
الف - 1) با توجه به شرط Realizability که بودی $\mathcal{H}$ داریم پس کم‌ترین متوسط خطا بر روی کل جمعیت از جمله هر Sample Set برابر با صفر خواهد بود. حال با توجه به این که فرضیه‌ای که A از یادگیری بر روی داده آموزشی داده است، دارای متوسط خطای صفر است و هم چنین چون در روش ERM، فرضیه‌ای بدست می‌آید که دارای کمترین خطا روی S باشد و با توجه به این که شرط Realizability برقرار است، این خطا برابر با صفر است پس خروجی الگوریتم A می‌تواند خروجی الگوریتم ERM باشد چون کمترین خطا h بر روی S که قابل دسترسی بوده است، داده است.

سوال T12 (ادامه)

الف-2)

$$\text{if } m \geq \frac{4\log\frac{4}{\delta}}{\epsilon} \xrightarrow[\text{least Prob. } 1-\delta]{\text{with at}} |L_D(h_A) - L_S(h_A)| \leq \epsilon.$$

$h_A : A$ خروجی الگوریتم



(1): باتوجه به اینکه متوسطی $R(S)$ برابر با صفر است پس هیچ نقطه ای را اشتباه دسته بندی نمی کند حال چون در هریک از R_i ها با احتمال $\frac{\epsilon}{4}$ ممکن است نقطه ای وجود داشته باشد که ما label آن را نمی دانیم پس احتمال خطا وجود دارد پس نمی تواند $R(S)$ مرز از R_i ها بزرگ پس در R^* قرار دارد.

$$\Rightarrow R(S) \subseteq R^*.$$

(2): حال اگر فرض کنیم که $R(S)$ شامل تمام نقاط R_i ها باشد، در بدترین حالت ممکن با حداکثر احتمال ϵ، نقطه ای وجود داشته باشد که negetive باشد و در این صورت اشتباه تشخیص داده می شود که در این صورت خطای ما برابر با خواهد بود با:

$$L_S(h_A) \leq \underbrace{1}_{\text{loss}} \times \underbrace{\epsilon}_{\text{prob.}} = \epsilon$$

(3):

$$\underset{S\sim D}{\text{Prob}}\left[\underbrace{R_i \not\subseteq S}_{A_i}\right] = \underset{\substack{x_j\sim D \\ j=1,2,\dots,m \\ x_j \in S}}{\text{Prob}}\left[x_j \notin R_i\right] = \left(1-\frac{\epsilon}{4}\right)^m$$

(4):

$$\text{Prob}\left\{\bigcup_{i=1}^{4} A_i\right\} \leq \sum_{i=1}^{4} \text{Prob}[A_i] \leq 4\left(1-\frac{\epsilon}{4}\right)^m \leq 4e^{-\frac{m\epsilon}{4}}.$$

ادامه سوال T12

پس خواهیم داشت:

$$Prob\left[L_D(h_S) > \varepsilon\right] \leq Prob\left\{\bigcup_{i=1}^{4} A_i\right\} \leq \underbrace{4e^{-\frac{m\epsilon}{4}}}_{\delta}$$

$$\rightarrow m \geq \frac{4\log\frac{4}{\delta}}{\epsilon}$$

ب) مقدار $m_{\mathcal{H}}(\varepsilon,\delta)$ بست آمده - با فرض اینکه شرط Realizability برقرار است، محاسبه شده و نرخ همگرایی خطا بهتری نسبت به $\frac{d+\log 1/\delta}{\epsilon^2}$ بست آورده ایم و آن را هیچ در (PAC learning)

بخشیده ایم.

سوال T13

$$l(h_S,(x,y)) = \begin{cases} 1 & h(x)\neq y \\ 0 & h(x)=y \end{cases}$$

$$\varepsilon_{bayes} = \underset{x}{E}\,\underset{y|x}{E}\left[l(h_S,(x,y))\right]$$

برای کفیه کردن مقدار ε_{bayes} - کافی است که برای هر x در $\mathcal{X}$، مقدار $E_{y|x}[l(h_S,(x,y))]$ را کفیه کنیم که در آن صورت مقدار متوسط آن نیز کفیه خواهد شد.

$$Prob\left[l(h_S,(x,y)) \mid X=x\right] = \begin{cases} \pi(x) & h(x)=0 \\ 1-\pi(x) & h(x)=1 \end{cases}$$

حال برای امید ریاضی این احتمال داریم:

ادامه سوال T13

$$\underset{y|x}{E}[l(h,(x,y))\,|\,x=x] = \begin{cases} \pi(x) & h(x)=0 \\ 1-\pi(x) & h(x)=1 \end{cases}$$

صفر

یک

پس متوسط خطای ما برای هر x یا $\pi(x)$ است و یا $1-\pi(x)$. حال کمترین متوسط خطا برابر است با کمترین مقدار بین دو عبارت $\pi(x)$ و $1-\pi(x)$، بنابراین برای هر $x \in X$، پس خواهیم داشت:

$$\forall x \in X: \quad \min \underset{y|x}{E}[l(h,(x,y))] = \min(\pi(x), 1-\pi(x))$$

از طرفی برای تخمین bayes داریم:

$$f(x) = \begin{cases} 1 & \pi(x) \geq \frac{1}{2} \\ 0 & \pi(x) < \frac{1}{2} \end{cases} \Rightarrow l(f_D,(x,y)) = \begin{cases} 1-\pi(x) & \pi(x) \geq \frac{1}{2} \\ \pi(x) & \pi(x) < \frac{1}{2} \end{cases}$$

$$\Rightarrow \underset{y|x}{E}[l(f_D,(x,y))] = \begin{cases} 1-\pi(x) & \pi(x) \geq \frac{1}{2} \\ \pi(x) & \pi(x) < \frac{1}{2} \end{cases}$$

$$= \min(\pi(x), 1-\pi(x)).$$

متوسط خطای تخمین‌گر bayes، دقیقاً برابر با کمترین خطایی که است که کل hها ممکن می‌توانند داشته باشند. $\Leftarrow$

$$L_D(f) = \varepsilon_{bayes} \leq L_D(h) \quad : \forall h$$

سوال T14

الف) با توجه به i.i.d بودن $l(h,(x_i,y_i))$ها برای hهای رندوم از Hoeffding thm استفاده می‌کنیم.

$$\text{Prob}[|L_D(h) - \underbrace{L_S(h)}_{\eta}| > \underbrace{\varepsilon}_{\eta}] \leq 2e^{-2m\epsilon^2}$$

اگر $\varepsilon = \eta$:

$$\Rightarrow \text{Prob}[L_D(h) > 2\eta] \leq 2e^{-2m\eta^2}$$

ب) از آن جایی که نمی دانیم که آیا الگوریتم A ، h_S را براساس $l(h_S(x_i), y_i)$ های دهد یا نه، یا به بیان دیگر این که h_S ، با توجه به مقادیر خطاها بدست می آید یا نه، نمی توانیم از Hoeffding استفاده کنیم چرا که ممکن است $l(h_S(x_i), y_i)$ بهم وابسته باشند یعنی

حال چون $|\mathcal{H}|$ محدود است پس UC داریم و می توانیم از bound آن استفاده کنیم:

$$\forall h \in \mathcal{H} : \mathbb{P}rob\left[|L_D(h) - L_S(h)| > \varepsilon\right] \leq 2|\mathcal{H}|\, e^{-2m\varepsilon^2}$$

حال با قرار دادن $L_S(h_S) = \gamma$ و $\varepsilon = \gamma$ داریم:

$$\mathbb{P}rob\left[|L_D(h) - \gamma| > \gamma\right] \leq 2|\mathcal{H}|\, e^{-2m\gamma^2}$$

$$\Rightarrow \mathbb{P}rob\left[L_D(h) > 2\gamma\right] \leq 2|\mathcal{H}|\, e^{-2m\gamma^2}.$$

ج)

$$\varepsilon_{bayes} = 3\eta \Rightarrow \forall h \in \mathcal{H} \text{ کل مجموعه ها ممکن} : L_D(h) \geqslant 3\eta.$$

پس خواهیم داشت که:

$$\mathbb{P}rob\left[L_D(h) > 2\eta\right] = 1$$

با توجه به ε_{bayes} ای که در این مسئله داده شده می دانیم که نمی تواند متوسط خطا برای هر h روی کل توزیع، کم تر از 3η باشد، حال احتمال اینکه برای h رندم دوستت الف متوسط خطا بر روی کل distribution بزرگ تر از 2η باشد 1 است. (بنابر هر D ممکن است.)

د) خیر، به نامساوی Hoeffding بر این اساس که ما اطلاعاتی در مورد توزیع داده نداریم (فرض است) و بر اساس آن نامساوی مورد نظر بدست می آید، حال اگر اطلاعاتی در مورد توزیع داشته باشیم قطعاً بر روی احتمالات ما تاثیر خواهد داشت و آن را بهبود می بخشد.