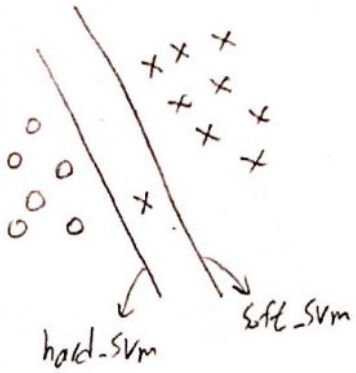
تمرین ۴ام یادگیری ماشین

آرمین سعادت
۹۶۱۰۵۸۲۹

سئوال ۱)

الف :

یک) خیر جواب سخت دوم لزوماً کم نیست. مثلاً داده‌های زیر را در نظر بگیرید.



در این داده‌ها مشخص است که خط مناسب همان است که توسط Soft-SVM مشخص شده و در واقع آن یک × به عنوان نویز باید در نظر گرفته شود.
اما در hard-SVM به همان یک نویز هم خیلی بها می‌دهد و عملاً همان یک × را به عنوان support-vector در نظر می‌گیرد.

اما اگر در Soft-SVM پارامتر C را ∞ بگذاریم، عملاً داریم مدل را مجبور می‌کنیم که misclassified کمتر یا حتی تراز کردن داده در margin خیلی بها دهد و اینگونه عملاً مانند همان hard-margin عمل می‌کند.

تابع خرد: $\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i$

اگر به ∞+ میل کند، مدل معادل hard-SVM می‌شود.

دو) برای Soft-SVM داریم:

$$\min_{w_1, w_2, \xi_i} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i$$

$$y^n(w^T x^n + w_0) \geq 1 - \xi_i$$

$\xi_i > 1$: داده‌ی i ام به اشتباه دسته‌بندی شده است. (misclassified)

$1 > \xi_i > 0$: داده i ام به درستی دسته‌بندی شده اما داخل margin قرار می‌گیرد.

$\xi_i = 0$: داده i ام با margin مناسب به درستی دسته بندی شده است.
در واقع یا داده i ام روی margin است یعنی خط بردار پشتیبان است.
یا این که کلاً inactive است. در هر ۲ صورت خارج margin (یا روی margin) و درست است.

$\lambda_i = 1$: این خیلی حالت خاصی است. با این مضاست که داده iام دقیقاً روی صفحه جداکننده قرار دارد.

به عبارتی $(w^T x_i + w_0) = 0$ شده است. و از آن جا که برای تابع Sgn داریم:

$$Sgn(x) \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}$$

عملاً داده iام به دسته خاصی تعلق می‌گیرد. مگر این که یک نرمِ اضافه کنیم که مثلاً $Sgn(x) \geq 0$ را به دسته‌ی ۱ اختصاص دهیم.

(سه)

$$\text{Primal problem: } p^* = \min_{x} \max_{\{\alpha_i \geq 0\}, \{\lambda_i\}} L(x, \alpha, \lambda).$$

$$\text{Dual problem: } d^* = \max_{\{\alpha_i \geq 0\}, \{\lambda_i\}} \min_{x} L(x, \alpha, \lambda)$$

$$\text{in general: } d^* \leq p^* \xrightarrow[h \text{ is affine}]{f \text{ and } g \text{ convex}} d^* = p^*$$

در حالت خاص برای SVM داریم:

$$\text{Primal: } \min_{w, w_0} \max_{\{\alpha_n \geq 0\}} \left\{ \frac{1}{2}\|w\|^2 + \sum_{n=1}^{N} \alpha_n \left(1 - y^{(n)}(w^T x^{(n)} + w_0)\right) \right\}$$

$$\text{dual: } \max_{\{\alpha_n \geq 0\}} \min_{w, w_0} \left\{ \frac{1}{2}\|w\|^2 + \sum_{n=1}^{N} \alpha_n \left(1 - y^{(n)}(w^T x^{(n)} + w_0)\right) \right\}$$

استفاده از دو دیدگاه مرجح است زیرا:

- معمولاً حل آن راحت‌تر است.
- دانش و بینش (insights) بیشتری در رابطه با نزدیک‌ترین صفحه جداکننده بهینه در اختیار ما می‌گذارد.
- این امکان را به ما می‌دهد که از روش‌های مبتنی بر Kernel بهره ببریم.

چهار)

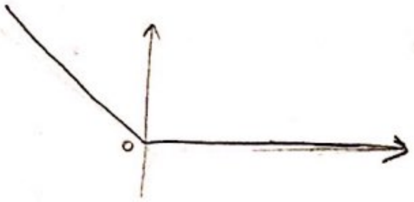
Perceptron : $\max\left\{-y^{(n)} h_\theta(x^{(n)}), 0\right\}$

Logistic regression: $-\left\{y^{(n)}\log\left(h_\theta(x^{(n)})\right) + (1-y^{(n)})\log\left(1-h_\theta(x^{(n)})\right)\right\}$ → log-loss or cross entropy

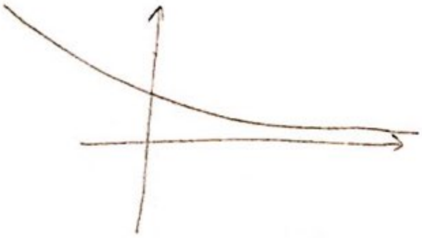
hard SVM : $\frac{1}{2}\|W\|^2$

soft SVM : $\frac{1}{2}\|W\|^2 + C\sum_{n=1}^{N}\underbrace{\max\left(0, 1-y^{(n)}(W^T x^{(n)} + w_0)\right)}_{\text{hinge loss}}$

Perceptron Loss :



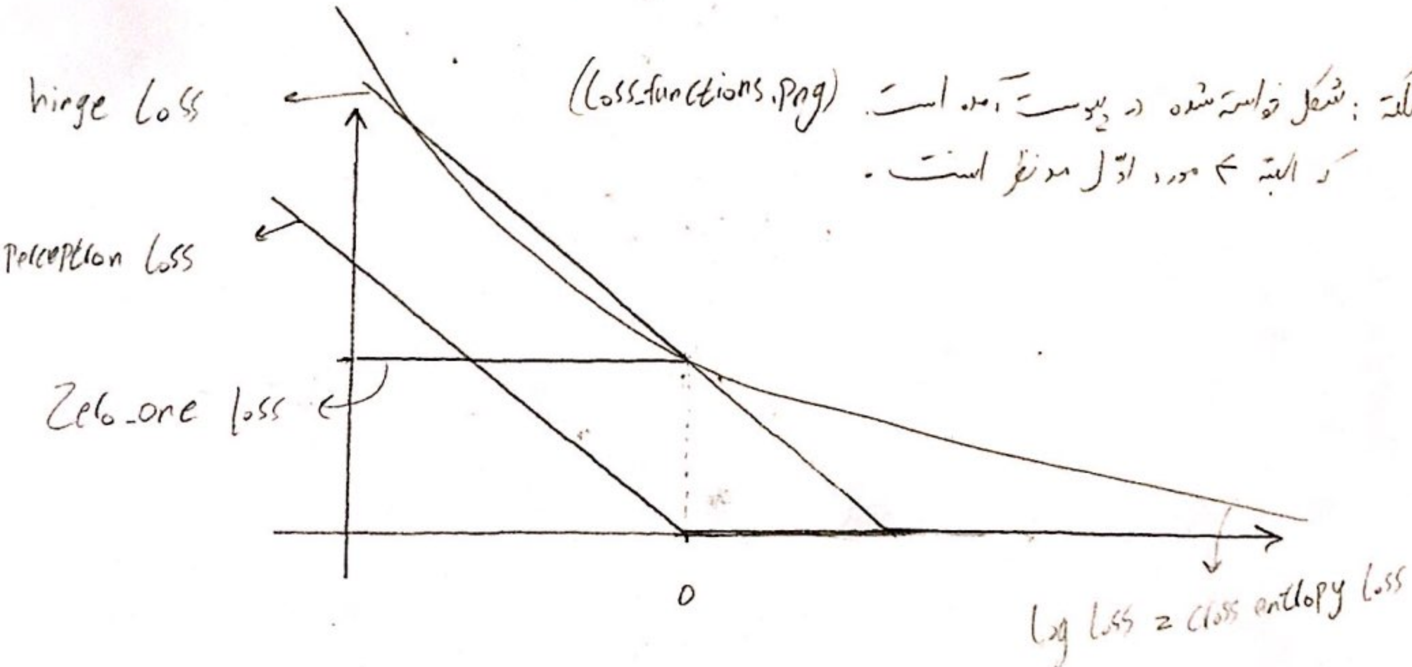
Logistic regression → cross entropy Loss :



Soft_SVM → hinge Loss :

نکته: شکل خواسته شده در پیوست آمده است. (Loss_functions.png)

که البته ← مورد اول مد نظر است.



hard_SVM → در واقع تابع هزینه ندارد و یک ترم Regularization هست.

یعنی $\|W\|^2$ باید کمینه شود.

نکته: شکل خواسته شده در پیوست آمده است.

ب) تحلیل تابع لولا:

$$\xi_i = \max\{0, 1 - y^{(n)}(w^T x^{(n)} + w_0)\}$$

$$\min \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N}\xi_n$$

بنابراین تنها تفاوت در تابع لولا $(\mathcal{L}(w, w_0, \xi, \alpha, \beta))$ در عبارت $C\sum_{n=1}^{N}\xi_n^2$ خواهد بود.

چرا که شرایط مسئله یعنی $\xi_n \geq 0$ ، $y^{(n)}(w^T x^{(n)} + w_0) \geq 1 - \xi_n$ مانند حالت قبل است. پس داریم:

$$\mathcal{L}(w, w_0, \xi, \alpha, \beta) = \frac{1}{2}\|w\|^2 + C\sum_{n=1}^{N}\xi_n^2 + \sum_{n=1}^{N}\alpha_n\left(1 - \xi_n - y^{(n)}(w^T x^{(n)} + w_0)\right) - \sum_{n=1}^{N}\beta_n\xi_n$$

پس مقادیر $\nabla_w \mathcal{L}$ ، $\frac{\partial \mathcal{L}}{\partial w_0}$ مانند حالت اصلی می‌شود.

اما برای مقدار $\frac{\partial \mathcal{L}}{\partial \xi_n}$ خواهیم داشت: (برای حالت اصل)

$$\frac{\partial \mathcal{L}}{\partial \xi_n} = C - \alpha_n - \beta_n = 0 \Rightarrow \alpha_n = C - \beta_n \,,\ \alpha_n \geq 0,\ \beta_n \geq 0 \Rightarrow 0 \leq \alpha_n \leq C$$

برای حالت جدید:

$$\frac{\partial \mathcal{L}}{\partial \xi_n} = 2C\xi_n - \alpha_n - \beta_n = 0 \Rightarrow \alpha_n = 2\xi_n \times C - \beta_n$$

$$\left.\begin{array}{l}\alpha_n \geq 0\\ \beta_n \geq 0\end{array}\right\} \quad 0 \leq \alpha_n \leq 2\xi_n \times C \Rightarrow C \geq 2C \longrightarrow 0 \leq \frac{\alpha_n}{\xi_n} \leq C$$

تحلیل تابع هزینه:

در این تابع جدید، به جای ξ_i ها از ξ_i^2 ها استفاده شده است. بنابراین به داده‌های نویز حساسیت بیشتری را نشان می‌دهد. چرا که برای داده‌های نویز، یعنی داده‌هایی که اشتباه دسته‌بندی شده‌اند، $\xi_i > 1$ است در نتیجه $\xi_i^2 > \xi_i$ و عملاً داده‌های نویز را دارد بیشتر از حالت عادی جریمه می‌کند پس باید انتظار داشته باشیم داده‌های نویز کمتر شوند. از طرفی داده‌های درون margin را کمتر جریمه می‌کند چرا که برای آن‌ها $0 < \xi_i < 1$ پس $\xi_i^2 < \xi_i$ خواهد بود.

پس اگر این ۲ حرف را کنار هم بگذاریم، این تابع جدید علاقه بیشتری دارد که margin را کوچک‌تر کند تا داده‌های کمتری نویزی شوند و ایراد کمتری بیند که margin کوچک شود و داده‌ها داخل margin بیفتند. یک جورایی می‌توان گفت شبیه SVM-Soft در حالت عادی عمل می‌کند با این فرق که C آن را زیاد کرده باشیم.

ب)

$$\mathcal{L}(w, w_0, \alpha) = \frac{1}{2}\|w\|^2 + \sum_{n=1}^{4} \alpha_n \left(1 - y^{(n)}(w^T x^{(n)} + w_0)\right)$$ یک)

$$\max_{\{\alpha_n \geq 0\}} \min_{w, w_0} \mathcal{L}(w, w_0, \alpha) \longrightarrow$$ رابطه‌ی بهینه سازی

بنابراین $N=4$ ، x_1 تا x_4 هر کدام در صورت سوال مشخص است. y_1 تا y_4 هم در صورت سوال مشخص است ($y_2 = y_3 = 1$ ، بقیه -1). پارامترهای w و w_0 نیز از رابطه‌ی زیر به دست می‌آیند:

$$\nabla_w \mathcal{L}(w, w_0, \alpha) = 0 \Rightarrow w = \sum_{n=1}^{4} \alpha_n y^{(n)} x^{(n)}$$

$$\frac{\partial \mathcal{L}(w, w_0, \alpha)}{\partial w_0} = 0 \Rightarrow \sum_{n=1}^{N} \alpha_n y^{(n)} = 0$$

با جایگذاری خواهیم داشت:

$$\mathcal{L}(\alpha) = \sum_{n=1}^{4} \alpha_n - \frac{1}{2} \sum_{n=1}^{4} \sum_{m=1}^{4} \alpha_n \alpha_m y^{(n)} y^{(m)} x^{(n)T} x^{(m)}$$

که به ازای $\alpha = \arg\max_{\alpha} \mathcal{L}(\alpha)$ به دست می‌آید.

پس از آن به ازای یک خط که $\alpha_n > 0$ دارد (بردار پشتیبان است) از رابطه‌ی زیر برای به دست آوردن w_0 استفاده می‌کنیم:

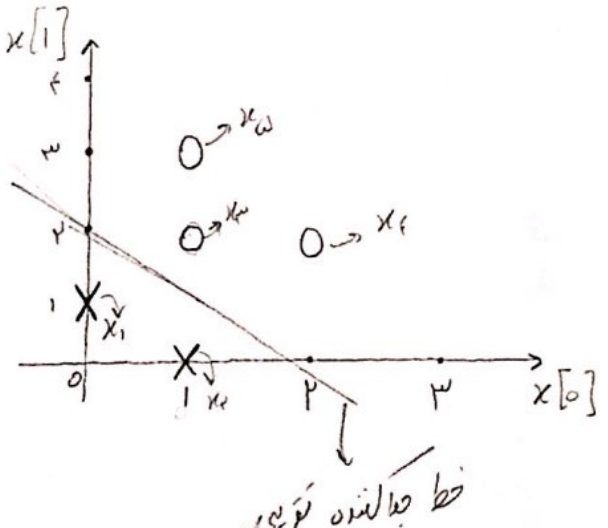
$$y^{(s)}(w^T x^{(s)} + w_0) = 1 \Rightarrow w_0 = y^{(s)} - w^T x^{(s)}$$

پس مشکل به دست آوردن α_n ها هست.

د)

$y^{(n)} = 1 \longrightarrow X$

$y^{(n)} = -1 \longrightarrow O$



بنابراین x_4 و x_3 نمی‌توانند بردار پشتیبان باشند.

(نکته: در واقع خط جدا کننده $x[1] = -x[0] + 2$ است اما این جا تقریبی رسم شده است.)

سه)

α_5 و α_4 صفرند. پس برای $\mathcal{L}(\alpha)$ داریم تنها یک بردار داریم:

$$\mathcal{L}(\alpha) = \sum_{n=1}^{3} \alpha_n - \frac{1}{2} \sum_{n=1}^{3} \sum_{m=1}^{3} \alpha_n \alpha_m y^{(n)} y^{(m)} x^{(n)T} x^{(m)}$$

$$= (\alpha_1+\alpha_2+\alpha_3) - \frac{1}{2} \left\{ \begin{array}{l} \alpha_1^2 \times 1 + \alpha_1\alpha_2 \times 0 + \alpha_1\alpha_3 \times -2 + \alpha_2\alpha_1 \times 0 + \alpha_2^2 \times 1 + \alpha_2\alpha_3 \times -1 \\ \alpha_3\alpha_1 \times -2 + \alpha_3\alpha_2 \times -1 + \alpha_3^2 \times 5 \end{array} \right\}$$

$$= \alpha_1+\alpha_2+\alpha_3 - \frac{1}{2} \left\{ \alpha_1^2 - 4\alpha_1\alpha_3 + \alpha_2^2 - 2\alpha_2\alpha_3 + 5\alpha_3^2 \right\}$$

از طرفی داریم:

$$\sum_{n=1}^{3} \alpha_n y^{(n)} = 0 \Rightarrow \boxed{\alpha_1 + \alpha_2 = \alpha_3}$$

با جایگذاری در رابطه $\mathcal{L}(\alpha)$ خواهیم داشت:

$$\mathcal{L}(\alpha) = 2\alpha_1 + 2\alpha_2 - \frac{1}{2} \left\{ \alpha_1^2 - 4\alpha_1 + 4\alpha_1\alpha_2 + \alpha_2^2 - 4\alpha_2 - 2\alpha_1\alpha_2 + 5\alpha_1^2 + 5\alpha_1\alpha_2 + 5\alpha_2^2 \right\}$$

$$= 2\alpha_1 + 2\alpha_2 - \alpha_1^2 - 2\alpha_1\alpha_2 - 2\alpha_2^2 = \mathcal{L}(\alpha_1, \alpha_2)$$

چهار) باید $\mathcal{L}(\alpha_1, \alpha_2)$ $\underset{\alpha_1,\alpha_2}{\arg\max}$ را به دست آوریم.

$$\left.\begin{array}{l} \dfrac{\partial \mathcal{L}(\alpha_1,\alpha_2)}{\partial \alpha_1} = 2 - 2\alpha_1 - 2\alpha_2 = 0 \\ \dfrac{\partial \mathcal{L}(\alpha_1,\alpha_2)}{\partial \alpha_2} = 2 - 2\alpha_1 - 4\alpha_2 = 0 \end{array}\right\} \begin{array}{l} \alpha_1 = 1 \\ \alpha_2 = 0 \end{array} \Big\} \alpha_3 = 1$$

پنج)

$$w = \sum_{n=1}^{5} \alpha_n y^{(n)} x^{(n)} \xrightarrow{\alpha_3=\alpha_4=\alpha_5=0} \alpha_1 y^{(1)} x^{(1)} + \alpha_2 y^{(2)} x^{(2)} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$$

$$w_0 = y^{(s)} - w^T x^{(s)} = -1 - \begin{bmatrix} -1 \\ -1 \end{bmatrix}^T \begin{bmatrix} 1 \\ 2 \end{bmatrix} = -1 - (-3) = 2$$

پس رابطه به شکل $w^T x + w_0$ که $w = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$, $w_0 = 2$.

$$\Rightarrow \begin{bmatrix} -1 \\ -1 \end{bmatrix}^T \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + 2 = \boxed{-x_1 - x_2 + 2} \longrightarrow \text{رابطه مرز}$$

$$\Rightarrow y^{(n)} = sgn\left\{ -x_1^{(n)} - x_2^{(n)} + 2 \right\}$$

مسئله ۲) هسته

الف) K_1 , K_2 هسته معتبر هستند. پس یک فضای بردای f_1 , f_2 وجود دارد که:

$$K_1(x, x') = f_1(x)^T f_1(x')$$
$$K_2(x, x') = f_2(x)^T f_2(x')$$

یک: کافی است فضای برداری f_3 را از کنار هم قرار دادن f_1 , f_2 تشکیل دهیم.

$$f_3(x) = \left[f_1(x)^T, f_2(x)^T \right]^T$$

$$\Rightarrow K_3(x, x') = f_3(x)^T f_3(x') = \left[f_1(x)^T, f_2(x)^T \right] \begin{bmatrix} f_1(x') \\ f_2(x') \end{bmatrix} = f_1(x)^T f_1(x') + f_2(x)^T f_2(x')$$

$$= K_1(x, x') + K_2(x, x') \longrightarrow$$

پس فضای f_3 وجود دارد که K_3 در آن ضرب داخلی باشد.
پس K_3 نیز یک هسته معتبر است.

دو:

$$K_4(x, x') = K_1(x, x') \cdot K_2(x, x') = f_1(x)^T f_1(x') \; f_2(x)^T f_2(x')$$

$$f_1(x) = \left[f_{11}(x), f_{12}(x), \ldots, f_{1n}(x) \right]^T , \quad f_2(x) = \left[f_{21}(x), f_{22}(x), \ldots, f_{2m}(x) \right]^T$$

$$\Rightarrow K_4(x, x') = \sum_{i=1}^{n} f_{1i}(x) f_{1i}(x') \times \sum_{j=1}^{m} f_{2j}(x) f_{2j}(x') = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(f_{1i}(x) f_{2j}(x) \right) \left(f_{1i}(x') f_{2j}(x') \right)$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{m} A_{ij}(x) A_{ij}(x') \;, \quad A_{ij}(x) = f_{1i}(x) f_{2j}(x) \Longrightarrow$$

برای K_4 یک فضای برداری مانند A وجود دارد که $K_4(x, x')$ در واقع ضرب داخلی x و x' در فضای A است.

مسئله ۲:

$a \geq 0 \Rightarrow a = (\sqrt{a})^2$

$aK_1(x,x') = a f_1(x)^T f_1(x') = \sqrt{a} f_1(x)^T \cdot \sqrt{a} f_1(x') = (\sqrt{a} f_1(x))^T (\sqrt{a} f_1(x'))$

$= f_3(x)^T f_3(x') = K_3(x,x') \quad , \quad f_3(x) = \sqrt{a} f_1(x)$

قسمت ۲:

$K_5 = \exp(K_1)$

با استفاده از بسط تیلور داریم:

$$\exp(K_1) = \exp(0) + \exp(0) K_1 + \frac{\exp(0)}{2!} K_1^2 + \frac{\exp(0)}{3!} K_1^3 + \dots$$

$$\Rightarrow \exp(K_1) = 1 + K_1 + \frac{1}{2} K_1^2 + \frac{1}{6} K_1^3 + \dots$$

$$\Rightarrow \exp(K_1) = \sum_{j=0}^{\infty} \exp(0) \times K_1^j \times \frac{1}{j!}$$

پس K_5 از جمع تعدادی عبارات به دست آمده است.

هر عبارت از ضرب یک عدد مثبت در حاصل ضرب تعدادی K_1 به دست آمده است.

طبق بخش ۵ و ۶ اثبات کردیم که ضرب ضریب مثبت در کرنل، باز هم یک کرنل می‌شود و ضرب ۲ کرنل نیز یک کرنل معتبر است.

پس هر عبارت یک کرنل معتبر است. پس جمع آن‌ها نیز طبق قسمت یک، کرنل معتبر است.

پس K_5 یک کرنل معتبر است.

نکته: عبارت اول $\exp(0) = 1$ است که خودش یک کرنل است.

اگر $K_4(x,x') = 1$ را در نظر بگیریم، فضای f_4 وجود دارد که:

$f_4(x) = [1]$

$\Rightarrow K_4(x,x') = f_4(x)^T f_4(x') = 1 \times 1 = 1 \Rightarrow K_4(x,x') = 1 \rightarrow$ کرنل معتبر است

ب) برای این که $K(A,B)$ یک کرنل باشد، باید فضای ویژگی را به گونه ای تعیین کنیم که در آن

$$Q(A)^T Q(B) = 2^{|A\cap B|}$$

تابع Q_U را اینگونه تعریف می کنیم:

$$Q_U(X) = \begin{cases} 1 & \text{if } U \subseteq X \\ 0 & \text{O.W.} \end{cases}$$

بنابراین ضرب داخلی را اینگونه تعریف می کنیم:

$$Q(A)^T Q(B) = \sum_{U \subseteq S} Q_U(A) Q_U(B)$$

سگمای بالا روی تمامی زیرمجموعه های S (که مجموعه مرجع است) اعمال می شود.

به ازای هر U، حاصل در صورتی ۱ می شود که هم $Q_U(A)=1$ هم $Q_U(B)=1$.

پس حاصل زمانی ۱ می شود که U هم زیرمجموعه‌ی A باشد هم زیرمجموعه‌ی B.

پس سگمای بالا به ازای $U \subseteq A\cap B$ برابر با ۱ است، در غیر این صورت ۰ است.

بنابراین حاصل عبارت بالا برابر تعداد زیرمجموعه های $A\cap B$ است که می شود $2^{|A\cap B|}$.

پس داریم:

$$Q(A)^T Q(B) = 2^{|A\cap B|} = K(A,B)$$

ب)

یک: اگر $x, x' \in R^2$ باشند خواهیم داشت:

$$K(x,x') = (c + x^T x')^2 = (c + x_1 x_1' + x_2 x_2')^2$$

$$= c^2 + 2c x_1 x_1' + 2c x_2 x_2' + x_1^2 x_1'^2 + x_2^2 x_2'^2 + 2 x_1 x_1' x_2 x_2'$$

$$\Rightarrow Q(x) = \left[c, \sqrt{2c}\, x_1, \sqrt{2c}\, x_2, x_1^2, x_2^2, \sqrt{2}\, x_1 x_2 \right]^T$$

حال برای $x, x' \in R^d$ داریم: (به صورت مشابه)

$$K(x,x') = (c + x^T x')^2 = (c + x_1 x_1' + x_2 x_2' + \dots + x_d x_d')^2$$

$$= c^2 + x_1^2 x_1'^2 + x_2^2 x_2'^2 + \dots + x_d^2 x_d'^2 + 2c x_1 x_1' + 2c x_2 x_2' + \dots + 2c x_d x_d' + 2 x_1 x_1' x_2 x_2' + \dots$$
$$+ 2 x_1 x_1' x_d x_d' + 2 x_2 x_2' x_3 x_3' + \dots + 2 x_{d-1} x_{d-1}' x_d x_d'$$

$$\Rightarrow Q(x) = \left[c, x_1^2, x_2^2, \dots, x_d^2, \sqrt{2c}\, x_1, \sqrt{2c}\, x_2, \dots, \sqrt{2c}\, x_d, \sqrt{2}\, x_1 x_2, \sqrt{2}\, x_1 x_3, \right.$$
$$\left. \dots, \sqrt{2}\, x_1 x_d, \sqrt{2}\, x_2 x_3, \dots, \sqrt{2}\, x_2 x_d, \dots, \sqrt{2}\, x_{d-1} x_d \right]^T$$

دو: عملاً $d+1$ بُعد آن صفر (یا به عبارت دیگر حذف) می‌شود.

$$K(x,x') = (x^T x')^2 = (x_1 x_1' + x_2 x_2' + \dots + x_d x_d')^2$$

$$\Rightarrow Q(x) = \left[x_1^2, x_2^2, \dots x_d^2, \sqrt{2}\, x_1 x_2, \sqrt{2}\, x_1 x_3, \dots \sqrt{2}\, x_1 x_d, \sqrt{2} x_2 x_3, \dots \sqrt{2} x_2 x_d, \dots \right.$$
$$\left. \dots, \sqrt{2}\, x_{d-1} x_d \right]^T$$

و جنبه دیگر بحثی bias نداریم.

سه :

$$K(x,x') = (c + x^T x')^M = (c + x_1x_1' + x_2x_2' + \dots + x_dx_d')^M$$

این عبارت تعدادی جمله ای می شود که هر کدام به تعداد متغیرهای x_i و x_i' دارند. بنابراین می توان به دو بخش x و x' تفکیک شود. (مانند قسمت (د)).

پس عملاً سوال این است که عبارت $(c + x_1x_1' + \dots + x_dx_d')^M$ چند عبارت می شود.

c را نیز به شکل x_0x_0' می بینیم. پس عملاً خواهیم داشت :

$$\left(\sum_{i=0}^{d} x_ix_i'\right)^M$$

برای سادگی x_ix_i' را معادل a_i در نظر می گیریم. پس داریم:

$$(a_0 + a_1 + a_2 + \dots + a_d)^M$$

هر عبارت به این شکل خواهد بود:

$$a_0^{b_0} \times a_1^{b_1} \times \dots \times a_d^{b_d}$$

$$0 \le b_i \le M$$

$$\sum_{i=0}^{d} b_i = M$$

پس مسئله تبدیل شد به این که $d+1$ متغیر b_i داریم که می خواهیم مجموعشان M شود، مقادیر صحیح بین 0 تا M می گیرند. حالات مختلف این b_i ها جواب مسئله است.

که برابر است با : $\binom{M+d}{d} = \binom{M+d}{M}$

پس تعداد ابعاد برابر است انتخاب M از $M+d$.

روش حل : K متغیر داریم که به ازای هرکدام :

$$\forall b_i : 0 \le b_i \,,\ \sum_{i=1}^{K} b_i = M$$

می توان به هر متغیر یک واحد اضافه کرد. پس مسئله تبدیل می شود به :

$$\forall b_i : 0 < b_i \,,\ \sum_{i=1}^{K} b_i = M+K$$

پس عملاً مانند این است که $M+K$ مهره داشته باشیم. پس $M+K-1$ جایگاه داریم بین مهره ها و کافی است $K-1$ جایگاه را انتخاب کنیم. که می شود $\binom{M+K-1}{K-1}$.

که در این مثال $K = d+1$ (چون c هم داریم). پس جواب نهایی می شود $\binom{M+d}{M} = \binom{M+d}{d}$

مسئله ۳) درخت تصمیم.

الف)

information gain → IG

$$IG(X_d, Y) = H(Y) - H(Y|X_d)$$

$$H(Y) = -\sum_{y_j \in Y} P(y_j) \log P(y_j)$$

$$H(Y|X_d) = -\sum_i \sum_j P(X_d = i, Y = j) \log P(Y = j | X = i)$$

$$P(Y = j) = \sum_{i \in X_d} P(Y = j, X_d = i) \Rightarrow H(Y) = -\sum_i \sum_j P(X_d = i, Y = j) \log P(Y = j)$$

$$\Rightarrow IG(X_d, Y) = -\sum_i \sum_j \left\{ P(X_d = i, Y = j) \times \left[\log P(Y = j) - \log P(Y = j | X_d = i) \right] \right\}$$

$$\log P(Y = j) - \log P(Y = j | X_d = i) = \log \frac{P(Y = j)}{P(Y = j | X_d = i)} = \log \frac{P(Y = j) P(X_d = i)}{P(Y = j, X_d = i)}$$

$$\Rightarrow IG(X_d, Y) = -\sum_i \sum_j P(X_d = i, Y = j) \log \frac{P(Y = j) P(X_d = i)}{P(Y = j, X_d = i)}$$

اگر X_d و Y از هم مستقل باشند، $P(Y = j) P(X_d = i) = P(Y = j, X_d = i)$ و این صورت عبارت داخل لگاریتم برابر ۱ می‌شود پس عبارت داخل سیگما به ازای هر زوج در $\log 1 = 0$ ضرب شده، $IG(X_d, Y) = 0$ خواهد بود.

از آن جا که IG نمی‌تواند منفی باشد، پس X_d در واقع کمترین gain را بین ویژگی‌ها دارد.
پس می‌تواند به عنوان ریشه انتخاب شود چرا که ID_3 ویژگی با بیشترین IG را در هر مرحله انتخاب می‌کند.
(مگر این که همه ویژگی‌ها مستقل از Y باشند!)

ب)

۲ دیدگاه کلی در رابطه با نحوه برخورد با شماره پرونده وجود دارد.

اگر به چشم یک متغیر پیوسته به آن نگاه کنیم و روی آن یک threshold قرار دهیم که مثلاً شماره‌های کمتر از K در دسته ۱ و مابقی در دسته ۲ قرار گیرند، در این صورت چون شماره پرونده مستقل از نوع بیماری است، به عنوان ویژگی انتخاب نخواهد شد. مگر این که بیمارستان سیاست خاصی در تعیین شماره پرونده‌ها داشته باشد. مثلاً یک حکاک ادیر انجام دهد و به بیماران بد خیم عدد کوچکتری نسبت به بیماران خوش خیم تخصیص دهد. در این صورت اگر K مناسب انتخاب شود، می‌تواند به عنوان ویژگی انتخاب شود و مدل هم خوب داده‌ها را دسته‌بندی کند و روی داده‌های تست هم خوب عمل خواهد کرد و overfitting نداریم. مثلاً فرض کنید بیمارستان به بیماران بدخیم شماره ۰ تا ۱۰۰۰، به بیماران خوش خیم شماره ۱۰۰۱ تا ۲۰۰۰ بدهد. اگر K = ۱۰۰۰ انتخاب شود، مدل بسیار مناسبی خواهیم داشت.

دیدگاه دوم این است که نحوه استفاده از شماره پرونده به شکل یک متغیر categorical باشد. یعنی مثلاً در ابتدایی ترین حالت، به ازای هر مقدار موجود برای شماره پرونده، یک دسته در نظر بگیریم. یعنی در نهایت تصمیم درخت بعد از عبور از این مرحله، به تعداد بیماران زننده خواهیم داشت که چون هر شماره منحصر به فرد است، در هر دسته دقیقاً یک بیمار قرار می‌گیرد و دقت روی داده‌های train ۱۰۰ درصد می‌شود. که در این صورت قطعاً شماره پرونده به عنوان ویژگی انتخاب خواهد شد. و خب این خیلی بد است چرا که مدل دقت ۱۰۰ درصدی دهد اما در حقیقت هیچ چیزی یاد نگرفته است و کاملاً overfit شده است. در نتیجه این روند تحلیل متغیر شماره پرونده مناسب نیست. پس به جای این که هر شماره پرونده را یک زننده در نظر بگیریم، می‌توانیم یک threshold قرار دهیم و یک بازه را به عنوان زننده در نظر بگیریم و در نهایت به نحوی تعداد زنندگان را محدود کنیم.

یا این که دقت را کامل بیاندازیم و پس از آن از بالا به پایین به کمک tune کردن آن بپردازیم.

اما در حالت کلی اگر شماره پرونده قبل از تشخیص بیماری به تیمار داده شود (یعنی عملاً مستقل باشد) و تعداد زنندگان و بازه‌های شماره پرونده را محدود در نظر بگیریم، تا حدودی IG خاصی نمی‌دهد و به عنوان ویژگی انتخاب نخواهد شد.

پس نکته مهم این است که به ازای هر شماره پرونده یک زننده نسازیم و تعداد دسته‌ها را محدود در نظر بگیریم.

ب) یک:

برای این کار کافی است الگوریتم ID3 را یک بار تا انتها اجرا کنیم. اثبات می شود که با شرایط گفته شده در صورت سوال، درخت تولید شده خطای آموزش صفر خواهد داشت.

یک برگ را در نظر بگیرید. اگر مقادیر این برگ یکسان باشند که مشکلی نداریم. چون یا همگی مقادیر درست را دارند، یا همگی یک مقدار غلط دارند که کافی است بچه آخرین ویژگی را برای این دسته به مقدار صحیح تغییر دهیم.

پس مشکل زمانی رخ می دهد که یک برگ مقادیر یکسان نداشته باشد. چون داریم یک برگ را بررسی می کنیم، پس ID3 متوقف شده و دیگر آن برگ را دسته بندی کرده است (چون اگر ویژگی دیگری نبود!).

شرط توقف ID3 این است که همه مقادیر یک دسته یکی باشند، یا این که از تمام ویژگی ها استفاده کرده باشد.

چون مقادیر این برگ یکسان نیستند و ID3 متوقف شده، پس نتیجه می گیریم که اعضای این دسته همگی از تمام ویژگی ها عبور کرده اند. چون درخت بدون تکرار است، از هر برگ به ریشه تنها یک مسیر وجود دارد. پس تمام اعضای یک دسته به ازای هر ویژگی، دقیقاً در یک دسته یکسان انتخاب شده اند. از آن جا که هر ویژگی، داده ها را به k دسته تقسیم می کند، پس ترکیبی در یک دسته به این معناست که داده ها ویژگی k ام یکسانی داشته اند.

پس تمام ویژگی های داده هایی که در یک دسته قرار دارند یکی است. پس عمداً 2^n یکسان داریم که n متفاوت دارند. اگر این را متناقض در نظر بگیریم، نتیجه می شود که در درخت مورد نظر، برگی وجود ندارد که مقادیر غیر یکسان داشته باشد پس خطای آموزش صفر است. اگر وجود ۲ داده یکسان که n متفاوت دارند را متناقض در نظر بگیریم، قطعاً به خطای آموزش صفر نخواهیم رسید چرا که تحت هیچ شرایطی آن ۲ داده یکسان لزوم قابل تشخیص نخواهند بود.

پس کافی است از الگوریتم ID3 استفاده کنیم.

یعنی به ازای ویژگی k ام، به ازای هر مقدار ممکن یعنی $\{1, 2, \dots, k\}$ یک نود شاخه تشکیل دهیم و تا زمانی که مقادیر موجود در یک دسته همگی یکسان نشده اند یا این که ویژگی ها تمام نشده است به کار ادامه دهیم.

طبق اثبات درخت حاصل خطای آموزش صفر خواهد داشت.

دو :

خیر.

مثال نقض: داده ها را در نظر می گیریم به صورتی که تنها یک ویژگی پیوسته وزن دارند. هدف تعیین جنسیت افراد است. مجموعه داده ما به این شکل است:

$$D = \{(۴۰, \text{پسر}), (۴۵, \text{دختر}), (۵۰, \text{پسر})\}$$

تنها یک ویژگی داریم ؛ درخت بدون تکرار است. پس در همان تقلیل اول کار تمام است.

فرض کنید آستانه را به نحوی انتخاب کنیم که خطای آموزش صفر شود. یعنی برای وزن های کمتر از K پسر باشد و...

$x < K \longrightarrow$ پسر $\Rightarrow$ $۴۰ < K$ ، $۵۰ < K \Rightarrow$ $۴۵ < K \longrightarrow$ دختر، پسر تشخیص داده شد

اگر وزن های کمتر از K را دختر بگیریم:

$x < K \longrightarrow$ دختر $\Rightarrow$ $۴۵ < K \Rightarrow ۴۰ < K$ ← پسر، دختر تشخیص داده شد.

پس چنین K ای وجود ندارد. پس هیچ درخت تصمیمی نمی توان ساخت.

سه :

وقتی در حالت گسسته براساس یک ویژگی تقسیم بندی را انجام می دهیم، مقدار دقیق آن ویژگی را می دانیم پس استفاده مجدد از آن ویژگی هیچ نفعی برای ما ندارد. پس استفاده مجدد از آن کمکی در افزایش دقت و ادامه دسته بندی به ما نمی کند.

اما در حالت پیوسته، وقتی براساس یک ویژگی تقسیم را انجام می دهیم، در واقع مقدار دقیق آن را نمی دانیم و صرفاً برای مقدار آن یک بازه در نظر گرفته ایم. پس استفاده مجدد از آن ویژگی می تواند سودمند باشد و به ما کمک کند که مقدار اصلی آن ویژگی را برای هر داده به صورت دقیق تری تخمین بزنیم. در نتیجه می توانیم با مقادیر دقیق تری داده ها را بهتر تقلیل دهیم و دسته بندی را ادامه دهیم و در نهایت خطای آموزش را کاهش دهیم.

اصلاً برای همین است که قسمت "بله" جواب داشت چون با استفاده از هر ویژگی فقط یک بار، در حالت گسسته تمام اطلاعات موجود را به دست می آوردیم. اما در حالت پیوسته این چنین نیست.

مسئله ۴. یادگیری جمعی

الف)

$$H_M(x) = \frac{1}{M}\sum_{m=1}^{M} h_m(x)$$

$$\Rightarrow (H_M(x) - h(x))^2 = \left(\frac{1}{M}\sum h_m(x) - h(x)\right)^2 = \frac{1}{M^2}\left(\sum h_m(x)\right)^2 + h(x)^2 - \frac{2h(x)}{M}\sum_{m=1}^{M} h_m(x)$$

$$\Rightarrow E_{com} = E_x\left[\frac{1}{M^2}\left(\sum_{m=1}^{M} h_m(x)\right)^2 + h(x)^2 - \frac{2h(x)}{M}\sum_{m=1}^{M} h_m(x)\right]$$

$$= \frac{1}{M^2}E_x\left[\left(\sum_{m=1}^{M} h_m(x)\right)^2\right] + E_x\left[h(x)^2\right] - \frac{2}{M}E_x\left[h(x)\sum_{m=1}^{M} h_m(x)\right] \rightarrow \text{(1) برابر}$$

$$E_{avg} = \frac{1}{M}\sum_{m=1}^{M} E_x\left[(h_m(x) - h(x))^2\right] = \frac{1}{M}E_x\left[\sum_{m=1}^{M}(h_m(x) - h(x))^2\right]$$

$$= \frac{1}{M}E_x\left[\sum_{m=1}^{M}\left(h_m(x)^2 + h(x)^2 - 2h(x)h_m(x)\right)\right] = \frac{1}{M}E_x\left[\sum_{m=1}^{M} h_m(x)^2 + \sum_{m=1}^{M} h(x)^2 - 2h(x)\sum_{m=1}^{M} h_m(x)\right]$$

$$E_x\left[\frac{1}{M}\times M\, h^2(x)\right] = E_x\left[h^2(x)\right]$$

$$= \frac{1}{M}E_x\left[\sum_{m=1}^{M} h_m^2(x)\right] + \frac{1}{M}E_x\left[\sum_{m=1}^{M} h^2(x)\right] - \frac{2}{M}E_x\left[h(x)\sum_{m=1}^{M} h_m(x)\right] \rightarrow \text{(2) برابر}$$

با مقایسه برابر (1) و (2)، نتیجه می‌شود که:

$$E_{com} - E_{avg} = \frac{1}{M^2}E_x\left[\left(\sum_{m=1}^{M} h_m(x)\right)^2\right] - \frac{1}{M}E_x\left[\sum_{m=1}^{M} h_m^2(x)\right]$$

پس باید اثبات کنیم که این عبارت کوچک‌تر مساوی ۰ است.

لم کوشی ـ شوارتز :

$$\left(\sum_{i=1}^{n} u_i v_i\right)^2 \le \left(\sum_{i=1}^{n} u_i^2\right)\left(\sum_{i=1}^{n} v_i^2\right)$$

با قرار دادن $n=M$ ، $u_i=1$ و $v_i = h_i(x)$ خواهیم داشت :

$$\left(\sum_{i=1}^{M} 1 \times h_i(x)\right)^2 \le \sum_{i=1}^{M} 1^2 \times \sum_{i=1}^{M} h_i^2(x) .$$

$$\Rightarrow \left(\sum_{m=1}^{M} h_m(x)\right)^2 \le M \sum_{m=1}^{M} h_m^2(x) \longrightarrow$$ ۲ طرف را در $\frac{1}{M^2}$ ضرب می‌کنیم.

$$\Rightarrow \frac{1}{M^2}\left(\sum_{m=1}^{M} h_m(x)\right)^2 \le \frac{1}{M} \sum_{m=1}^{M} h_m^2(x) \longrightarrow$$ از دو طرف امید ریاضی می‌گیریم.

$$\Rightarrow \frac{1}{M^2} E_x\left[\left(\sum_{m=1}^{M} h_m(x)\right)^2\right] \le \frac{1}{M} E_x\left[\sum_{m=1}^{M} h_m^2(x)\right]$$

$$\Rightarrow \frac{1}{M^2} E_x\left[\left(\sum_{m=1}^{M} h_m(x)\right)^2\right] - \frac{1}{M} E_x\left[\sum_{m=1}^{M} h_m^2(x)\right] \le 0$$

$$\Rightarrow E_{Com} - E_{avg} \le 0 \Rightarrow \boxed{E_{com} \le E_{avg}}$$

ب)

$$E_{Com} = E_x\left[\left(\frac{1}{M}\sum_{m=1}^{M} h_m(x) - h(x)\right)^2\right] = E_x\left[\left(\frac{1}{M}\sum_{m=1}^{M} h_m(x) - h(x)\right)\left(\frac{1}{M}\sum_{l=1}^{M} h_l(x) - h(x)\right)\right]$$

$$= \frac{1}{M^2}\sum_{m=1}^{M}\sum_{l=1}^{M} E_x\left[\left(h_m(x) - h(x)\right)\left(h_l(x) - h(x)\right)\right]$$

حاصل امید ریاضی باقا ضرب $m \neq l$ برابر با صفر است. (طبق فرض مسئله).

پس عبارت بالا برابر است با:

$$\frac{1}{M^2}\sum_{m=1}^{M} E_x\left[\left(h_m(x) - h(x)\right)\left(h_m(x) - h(x)\right)\right]$$

$$= \frac{1}{M^2}\sum_{m=1}^{M} E_x\left[\left(h_m(x) - h(x)\right)^2\right] = \frac{1}{M} \times \left\{\frac{1}{M}\sum_{m=1}^{M} E_x\left[\left(h_m(x) - h(x)\right)^2\right]\right\}$$

$$= \frac{1}{M} E_{avg} \Rightarrow \boxed{E_{Com} = \frac{1}{M} E_{avg}}$$

مسئله ۵. (AdaBoost)

الف) فرض می‌کنیم $h_t = h_{t+1}$.

برای مرحله t داریم:

$$\epsilon_t = \frac{\sum_{i=1}^{N} w_t^{(i)} \times I(y^{(i)} \neq h_t(x^{(i)}))}{\sum_{i=1}^{N} w_t^{(i)}}$$

برای راحتی، $w_t^{(i)} \times I(y^{(i)} \neq h_t(x^{(i)}))$ را با $w_t^{(i)}$ نشان می‌دهیم.

برای راحتی، $w_t^{(i)} \times I(y^{(i)} = h_t(x^{(i)}))$ را با $w_t^{'(i)}$ نشان می‌دهیم.

پس w_t، وزن داده‌هایی است که h_t توانسته تشخیص بدهد، w'_t وزن داده‌هایی است که h_t به اشتباه تشخیص داده است. پس داریم:

$$\epsilon_t = \frac{\sum_{y^i \neq h_t(x^i)} w_t^{(i)}}{\sum_{y^{(i)} \neq h_t(x^i)} w_t^{(i)} + \sum_{y^i = h_t(x^i)} w_t^{'(j)}} < \frac{1}{2}$$

در Adaboost فقط وزن داده‌هایی که توسط h_t اشتباه تشخیص داده شده‌اند کم می‌شود.

پس وزن $w_{t+1}^{'(i)} = w_t^{'(i)}$ اما وزن داده‌های غلط تشخیص داده شده می‌شود:

$$w_{t+1}^{(i)} = w_t^{(i)} e^{\alpha_t}, \quad \alpha_t = \ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right) \Rightarrow \begin{cases} w_{t+1}^{(i)} = w_t^{(i)} \times \frac{1-\epsilon_t}{\epsilon_t} \\ w_{t+1}^{'(i)} = w_t^{'(i)} \end{cases}$$

دقت شود که داده‌هایی که با h_t درست تشخیص داده می‌شوند، دقیقاً همان داده‌هایی هستند که توسط h_{t+1} درست داده می‌شوند چرا که فرض کردیم $h_{t+1} = h_t$.

حال خطای ϵ_{t+1} را حساب می‌کنیم:

$$\epsilon_{t+1} = \frac{\sum w_{t+1}^{(i)}}{\sum w_{t+1}^{(i)} + \sum w_{t+1}^{'(j)}} = \frac{\frac{1-\epsilon_t}{\epsilon_t} \times \sum w_t^{(i)}}{\frac{1-\epsilon_t}{\epsilon_t} \sum w_t^{(i)} + \sum w_t^{'(j)}}$$

$$\frac{1-\epsilon_t}{\epsilon_t} = \frac{1}{\epsilon_t} - 1 = \frac{\sum w_t^{(i)} + \sum w_t^{(j)}}{\sum w_t^{(i)}} - 1 = \frac{\sum w_t^{(j)}}{\sum w_t^{(i)}}$$

$$\Rightarrow \epsilon_{t+1} = \frac{\frac{\sum w_t^{(j)}}{\sum w_t^{(i)}} \times \sum w_t^{(i)}}{\frac{\sum w_t^{(j)}}{\sum w_t^{(i)}} \times \sum w_t^{(i)} + \sum w_t^{(j)}} = \frac{\sum w_t^{(j)}}{2\sum w_t^{(j)}} = \frac{1}{2}$$

پس اگر $h_{t+1} = h_t$ باشد، ϵ_{t+1} دقیقاً برابر با ۰٫۵ می‌شود. پس خطای یادگیری ۰٫۵ است.
این خلاف فرض است که خطای یادگیرهای ضعیف از ۰٫۵ اکیداً کمتر است.

پس به تناقض رسیدیم. پس $h_{t+1} \neq h_t$

ب)

الگوریتم AdaBoost از تابع خطای نمایی استفاده می‌کند. فرض کنید $t-1$ یادگیر ساده به H_t اضافه شده باشد. یادگیر t ام را می‌خواهیم اضافه کنیم.

$$H_t(x) = \frac{1}{2}(\alpha_1 h_1(x) + \alpha_2 h_2(x) + \dots + \alpha_t h_t(x)) = H_{t-1}(x) + \frac{1}{2}\alpha_t h_t(x)$$

$$E = \sum_{i=1}^{N} e^{-y^{(i)} H_t(x^{(i)})} = \sum_{i=1}^{N} e^{-y^{(i)}\left[H_{t-1}(x^i) + \frac{1}{2}\alpha_t h_t(x^i)\right]}$$

$$= \sum_{i=1}^{N} \underbrace{e^{-y^{(i)} H_{t-1}(x^i)}}_{\text{در مرحله } t \text{ ثابت است و برابر با } w_t^{(i)} \text{ در نظر گرفته می‌شود.}} e^{-\frac{1}{2}\alpha_t y^{(i)} h_t(x^{(i)})} = \sum_{i=1}^{N} w_t^i \underbrace{e^{-\frac{1}{2}\alpha_t y^i h_t(x^i)}}_{\text{این عبارت باید کمینه شود}}$$

برای به دست آوردن α_t و h_t.

برای به دست آوردن h_t خواهیم داشت:

$$E = \sum_{i=1}^{N} w_t^i e^{-\frac{1}{2}\alpha_t y^i h_t(x^i)} = e^{\frac{-\alpha_t}{2}} \sum_{y^{(i)}=h_t(x^i)} w_t^i + e^{\frac{\alpha_t}{2}} \sum_{y^{(i)}\neq h_t(x^i)} w_t^i$$

$$= \underbrace{\left(e^{\frac{\alpha_t}{2}} - e^{\frac{-\alpha_t}{2}}\right)}_{\text{به مقدار } h_t \text{ ربطی ندارد. و در ضمن مقداری مثبت است.}} \sum_{y^{(i)}\neq h_t(x^i)} w_t^{(i)} + \underbrace{e^{\frac{-\alpha_t}{2}} \sum_{i=1}^{N} w_t^{(i)}}_{\text{مستقل از } h_t(x) \text{ است.}}$$

پس عملاً تابعی که باید بهینه شود به این شکل است:

$$\sum_{y^{(i)}\neq h_t(x^i)} w_t^{(i)} \Rightarrow J_t = \sum_{i=1}^{N} w_t^{(i)} \times I(y^{(i)} \neq h_t(x^{(i)}))$$

پس عملاً تعیین h_t ، نتیجه الگوریتم فقط به وزن داده هایی ربط دارد که به اشتباه دسته بندی شده اند.

حال به سراغ به روز رسانی وزن های بعدی.

طبق روش AdaBoost می دانیم به روز رسانی اینچنین اتفاق می افتد:

$$\left.\begin{aligned} w_{t+1}^{(i)} &= w_t^{(i)} e^{-\frac{1}{2}\alpha_t y^{(i)} h_t(x^{(i)})} \\ y^{(i)} h_t(x^{(i)}) &= 1 - 2\, I(y^{(i)} \neq h_t(x^{(i)})) \end{aligned}\right\} \Rightarrow w_{t+1}^{(i)} = w_t^{(i)} \underbrace{e^{-\frac{1}{2}\alpha_t}}_{\text{این قسمت مستقل از } i \text{ هست و عملا در همه وزن ها ضرب می شود پس تاثیری ندارد.}} e^{\alpha_t I(y^{(i)} \neq h_t(x^{(i)}))}$$

$$\Rightarrow w_{t+1}^{(i)} = w_t^{(i)} e^{\alpha_t I(y^{(i)} \neq h_t(x^{(i)}))}$$

پس داده هایی که به درستی تشخیص داده شده اند، I مقدار صفر خروجی می دهد، $e^0 = 1$ پس $w_{t+1}^{(i)} = w_t^{(i)}$

پس فقط داده هایی که اشتباه دسته بندی شده اند به روز رسانی می شوند.

وزن