۱۰.۱ با الگوریتم A، می‌دانیم:

$$P\left(L_D(A(S)) \le \min_{h\in H} L_D(h)+\epsilon\right) \ge 1-\delta$$

حال باید نشان دهیم که شرط A برقرار باشد و H را بتوان agnostic pac یاد گرفت

$$m_H(\epsilon,\delta) \le k\, m_H(\epsilon) + \left\lceil \frac{2\log(4k/\delta)}{\epsilon^2} \right\rceil$$

که $k=\lceil \log(\delta)/\log(\delta_0)\rceil$

فرض کنیم $\epsilon,\delta\in(0,1)$، k تا به اندازه‌ی $m_H(\epsilon/2)$ برداریم و الگوریتم

A را روی هر کدام اجرا کنیم داریم $\hat h_1, \hat h_2, \ldots, \hat h_k$.

توجه کنیم که احتمال $\min_{i\in[k]} L_D(\hat h_i) \le \min L_D(h)+\frac{\epsilon}{2}$، حداقل $1-\delta_0^k \ge 1-\delta/2$

باشد. پس ERM روی $\hat H=\{\hat h_1,\hat h_2,\ldots,\hat h_k\}$ اعمال کنیم که

training data هم به اندازه‌ی آخرین مقدار برابر با $\left\lceil \frac{2\log(4k/\delta)}{\epsilon^2} \right\rceil$ است و تابع خروجی

را $\hat h$ فرض کنیم. با استفاده از قضیه‌ی قبل می‌توانیم بگوییم با احتمال

حداقل $1-\delta/2$، $L_D(\hat h) \le \min_{i\in[k]} L_D(\hat h_i)+\frac{\epsilon}{2}$ است و هم چنین

$$L_D(\hat h) \le \min_{i\in[k]} L_D(\hat h_i)+\frac{\epsilon}{2} \le \min_{h\in H} L_D(h)+\epsilon$$

۱۶ . ۲

ثابت کنیم بیشتر از $O(dT\log(dT))$ که $d = VCdim(B)$ ، می‌خواهیم ثابت کنیم

با این تعریف منطقی می‌دانیم هر آنچه که برای B صادق خواهد بود.

درسته؟ آیا! $VCdim(B) \le VCdim(L(B,T))$ ، یک کلاس B بیابید که

$VCdim(B) = VCdim(L(B,T))$

برای هر $T \ge 1$.

— X را یک مجموعه‌ی متناهی از اندازه n بگیرید و B مجموعه‌ی توابع از

$X \to \{\pm 1\}$ باشد پس

برای هر $T \ge 1$

$VCdim(B) = VCdim(L(B,T)) = \log 2^n = n$

۲. $B = \{h_{j,b,\theta} : j \in [d], b \in \{-1,1\}, \theta \in \mathbb{R}\}$

به طوری که $h_{j,b,\theta}(x) = b \cdot sign(\theta - x_j)$

برای هر $j \in [d]$ در نظر بگیریم $B_j = \{h_{b,\theta} : b \in \{-1,1\}, \theta \in \mathbb{R}\}$ که

$h_{b,\theta}(x) = b \cdot sign(\theta - x_j)$

و $VCdim(B_j) = 2$

واضح است که $B = \bigcup_{j=1}^{d} B_j$ و هم چنین می‌دانیم

$$VCdim(B) \leq 16 + 2\log d$$

$2^{\lfloor \log d \rfloor}$

فرض کنید که $d = 2^k$ برای هر $K \in N$ و

بگیرید $A \in R^{k \times d}$ یک ماتریس به صورتی که ستون‌هایش از مجموعه $\{0,1\}^k$ باشد.

برای هر $i \in [k]$ بگیرید $x_i = A_i$. می‌توانیم مجموعه C را

$C = \{x_1, \dots, x_k\}$ تعریف کنیم. بگیرید $I \subseteq [k]$ ، می‌توانیم I رو با برچسب مثبت بزنیم وقتی که $[k]/I$

ستون j ام ماتریس منحصربفرد است و j را جوری داریم که $A_{ij} = x_{ij} = 1$ iff, $i \in I$

پس $h_{j,-1,\frac{1}{2}}(x_i) = 1$ iff $i \in I$

۴- فرض کنید $T>1$ عدد صحیحی باشد. ثابت کنید $VCdim(L(B_d,T)) \geq 0.5\, T\log(d)$

با توجه به راهنمایی: برای هر $i \in [Tk/2]$ فرض کنید $x_i = [i/k]_{A_{k\geq 0}}$. را در نظر می‌گیریم

مجموعه‌ی $C = \{x_i : i \in [Tk/2]\}$ توسط $L(B_d,T)$ ، shatter می‌شود

فرض کنید $I \subseteq [Tk/2]$ پس $I = I_1 \cup \dots \cup I_{T/2}$ که هر I_t

زیرمجموعه‌ی $\{(t-1)k+1, \dots, tk\}$ است. برای هر $t \in [T/2]$

فرض کنید j_t ستونی از A ($i.e.$ $A_{ij}=1$ iff $(t-1)k+i \in I_t$)

تعریف: $h(x) = \operatorname{sign}\Big(\big(h_{j_1,-1,\frac{1}{2}} + h_{j_1,1,\frac{3}{2}} + h_{j_2,-1,\frac{3}{2}} +$

$h_{j_2,1,\frac{5}{2}} + \dots + h_{j_{T/2-1},-1,T/2-\frac{3}{2}} + h_{j_{T/2-1},1,T/2-\frac{1}{2}}$

$+ h_{j_{T/2},-1,\frac{T}{2}-\frac{1}{2}}\big)(x)\Big)$

پس $h(x_i) = 1$ iff $i \in I$

پس $h \in L(B_d,T)$

11.1 فرض کنید توزیعی داریم که برچسب به صورت تصادفی با $P(y=1)=P(y=0)=\frac{1}{2}$

$h(x)=1$ اگر لیبل ۱ باشد و در غیر این صورت $h(x)=0$ می‌دهد. ثابت کنید تفاوت

estimate

بین leave-one-out و خطای حقیقی در چنین حالتی همیشه $\frac{1}{2}$ است.

که روی یک مجموعه داده آموزش ببیند و h تابعی باشد که الگوریتم به ما داده.

توجه کنید که $L_D(h)=\frac{1}{2}$ (استقلال) چون که h تابع ثابت است

فرض کنید که پاریتی S را ۱ باشد. قرار دهید چند تا fold $(x,y) \in S$ ، بین دو حالت

تمایز قائل شویم:

۱. پاریتی $S/\{x\}$ مساوی ۱ . این نشان می‌دهد که $y=0$. وقتی از $S/\{x\}$

estimate

استفاده می‌کنیم خروجی الگوریتم تخمین می‌زند $h(x)=1$. پس leave-one-out از این fold ها

استفاده کند میشه ۱.

۲. پاریتی $S\setminus\{x\}$ برابر ۰ . اونوقت $y=1$ ، پس خروجی الگوریتم حدس می‌زند که $h(x)=0$

است که تابع leave-one-out estimate ، استفاده می‌کند از این fold ها هست ۱.

به طور مشابه بررسی fold ها . تخمین خطای h ، ۱ می‌باشد. پس تفاوت

بین estimate و خطای واقعی ، $\frac{1}{2}$ است. وقتی پاریتی S ، ۰ است هم به طور مشابه

این هم برای آن آنالیز می‌شود

۱۸.۲ :

$X = \{0,1\}^3$, $Y = \{0,1\}$

$((1,1,1),1)$

$((1,0,0),1)$

$((1,1,0),0)$

$((0,0,1),0)$

فرض کنید می خواهیم از این training set برای ساخت یک درخت تصمیم با عمق ۲ استفاده کنیم.

۱- فرض کنیم الگوریتم ID_3 را به عمق ۲ اجرا کنیم. فرض کنید در هر مرحله از آن برای اندازه‌گیری کیفیت هر ویژگی استفاده شده بر اساس آنتروپی باشد. و الگوریتم را برای ایجاد یک لیبل کند، یعنی از آنها بطور یکسان استفاده می کند. نشان دهید خطای آموزش درخت تصمیم حداقل $\frac{1}{4}$ است.

۲- پیدا کنید یک درخت تصمیم به عمق ۲ که خطای آموزش به صفر برسد.

- به عنوان مثال، می توانیم دو دسته فرضیه $H_1 \subseteq H_2 \subseteq \dots \subseteq H_k$, $i \in k$, $|H_i| = 2^i$

learning H_k is th agnostic pac model

که از کران زیر برای تابع ERM از h به دست می آید

$$L_D(h) \le \min_{h \in H_k} L_D(h) + \sqrt{\frac{2(k+1) + \log 1/\delta}{m}}$$

از طرف دیگر می‌توانیم از مثلی که در ادامه توضیح بدهیم استفاده کنیم

فرض کنید که j minimal index که شامل $h^* \in \arg\min_{h \in H} L_D(h)$ می‌باشد

است. قرار دهید که $r \in [k]$. با احتمال حداقل $1 - \delta/2k$

داریم
$$|L_D(\hat{h}_r) - L_V(\hat{h}_r)| \le \sqrt{\frac{1}{2\alpha m} \log \frac{4}{\delta}}$$

با استفاده از union bound می‌توانیم ببینیم که با احتمال حداقل $1 - \delta/2$ داریم

$$L_D(\hat{h}) \le L_V(\hat{h}) + \sqrt{\frac{1}{2\alpha m} \log \frac{4k}{\delta}}$$
$$\le L_V(\hat{h}_r) + \sqrt{\frac{1}{2\alpha m} \log \frac{4k}{\delta}}$$
$$\le L_D(\hat{h}_r) + 2\sqrt{\frac{1}{2\alpha m} \log \frac{4k}{\delta}}$$
$$= L_D(\hat{h}_r) + \sqrt{\frac{2}{\alpha m} \log \frac{4k}{\delta}}$$

$$L_D(\hat{h}) \le L_D(\hat{h}_j) + \sqrt{\frac{2}{\alpha m} \log \frac{4k}{\delta}}$$

$$L_D(\hat{h}_j) \le L_D(h^*) + \sqrt{\frac{2}{(1-\alpha) m} \log \frac{4|H_j|}{\delta}}$$
$$= L_D(h^*) + \sqrt{\frac{2}{(1-\alpha) m} \log \frac{4|H_j|}{\delta}}$$

با ادغام دو تا نامساوی مشاهده خواهیم کرد که با احتمال حداقل $1-\delta$ داریم

$$L_D(\hat{h}) \le L_D(h^*) + \sqrt{\frac{2}{\alpha m}\log\frac{4k}{\delta}} + \sqrt{\frac{2}{(1-\alpha)m}\log\frac{4}{}}$$

نتیجه می‌گیریم که

$$L_D(\hat{h}) \le L_D(h^*) + \sqrt{\frac{2}{\alpha m}\log\frac{4k}{\delta}} + \sqrt{\frac{2}{(1-\alpha)m}\left(j + \log\frac{4}{\delta}\right)}$$

با مقایسه این دو تا محدوده، می‌بینیم که "optimality" از S به طور قابل توجهی کوچک‌تر از k است و کران به دست آمده با استفاده از اکسپوننشیال کردن است. آنگاه اگر j موجب k لگاریتمی باشد، به طور دقیق‌تر، به بهبود لگاریتمی دست می‌یابیم.