۳.۲:

۱

اگر x منفی باشد

$$h_z(x) = \begin{cases} 1 & \text{if } x = z \\ 0 & \text{if } x \neq z \end{cases}$$

و اگر هیچ عضو منفی نداشته باشه $h^-(x) = 0 \quad \forall x \in X$

ما در الگوریتم می‌نویسیم اگر $x<0$ بود پس $h_x(x)$ رو برگردونه و اگر $x>0$ بود $h^-(x)=0$ رو برگردونه و این یک الگوریتم ERM است، زیرا در ERM فقط میره تابعی رو انتخاب می‌کنه که کم‌ترین خطای تجربی رو داشته باشه و در این‌جا هم تابع فوق یک عدد حتما صفر می‌گیره پس خطای تجربی‌مون برابر صفر می‌شه و واضح است که این الگوریتم یک ERM است.

۲ نشان دهیم PAC learnable هست و یک کران بالا براش معرفی کنیم

باید نشون بدیم برای هر $\epsilon, \delta \in (0,1)$ و توزیع D روی X

یک تابع $m_H : (\epsilon, \delta)^2 \to N$ هست که خطای حقیقی‌مون با احتمال $1-\delta$ از ϵ کمتره. اگر ما از h^- استفاده کنیم خیلی کامل هست.

حالا اگر بخواد $h_z(x)$ استفاده کنه فرض کنیم یه S رو انتخاب بکنیم این S دقیقا

باید ببینیم تابع $m_H : (0,1)^2 \to \mathbb{N}$ وجود دارد به طوریکه خطای حقیقی‌مون با احتمال $1-\delta$، از ϵ کمتر شود

$$P(L_{D,f}(h) \le \epsilon) = 1-\delta$$

اگر کلاً هیچ x مثبتی نباشد پس از h^- استفاده می‌کنیم که اینجا

$P(L_{D,f}(h) \le \epsilon) = 1$ خواهد بود پس ✓

اگر در خارج از S یک x_+ نباشد پس از $h_{x_+}(x)$ استفاده می‌کنیم بنابراین

باز هم $P(L_{D,f}(h_x(x)) \le \epsilon) = 1$ ✓

حال اگر در خارج S، x_+ با احتمال کمتر از ϵ بود پس $P(L_{D,f}(h(x)) \le \epsilon) = 1$

خواهد شد ✓

حال فرض کنیم که در S، x_+ نباشد ولی خارج آن x_+ با احتمال بیشتر از ϵ وجود دارد پس احتمال

انتخاب نشدن $1-\epsilon$ است که مطلوب ما است پس توی این m تا عضوی که می‌خواهیم

که با F نشون میدیم

برداریم احتمال انتخاب نشدن x_+، $(1-\epsilon)^m$ است پس:

$$P(F) \le (1-\epsilon)^m \le e^{-m\epsilon} < \delta$$

پس اگر $m_H \ge \left\lceil \frac{\log \frac{1}{\delta}}{\epsilon} \right\rceil$ باشد می‌توانیم بگوییم با احتمال $1-\delta$، خطا حداکثر ϵ می‌باشد.

$\mathcal{X} = \mathbb{R}^2 \quad \mathcal{Y} = \{0, 1\}$

۲) H مدل دایره‌های هم‌مرکز در صفحه باشند

$\mathcal{H} = \{h_r : r \in \mathbb{R}_+\}$

که $h_r(x) = \mathbb{1}_{[\|x\| \leq r]}$

نشان دهید H، PAC learnable است و

$m_{\mathcal{H}}(\epsilon, \delta) \leq \left\lceil \frac{\log(1/\delta)}{\epsilon} \right\rceil$

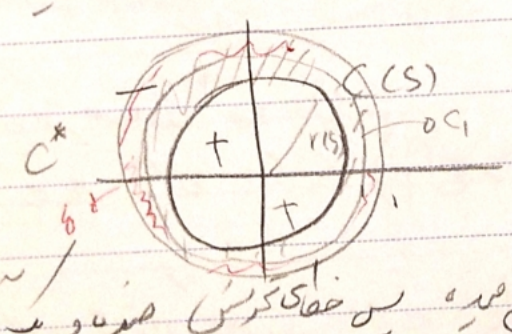
برای اثبات PAC learnable بودن باید نشان دهیم که تابع h با m_H داریم

به طوری که:

$P(L_{\mathcal{D},f}(h) \leq \epsilon) = 1 - \delta$

فرض کنیم الگوریتم A یک دایره‌ی C(S) دور بگیرد درونه با شعاع r(S) و هرچه داخلش لیبل + ،

بیرونش لیبل منفی می‌گیرند



پس نشان دهنده اینه که تونسته داده‌های S رو به درستی لیبل بده پس خطای آموزش صفر و یک الگوریتم

ERM است ،

فرض کنیم f ، دایره‌ی C^* رو لیبل بدون میده ،

$L_{\mathcal{D},f} = P(x \in \mathcal{X} : h(x) \neq f(x)) = P(C^* \setminus C(S))$

Subject: $r_1 < r^*$

فرض میکنیم: $C_1 = \{x \in \mathbb{R}^2 , r_1 \le \|x\| \le r^*\}$

برای همه حالت ها اولین فضا اون قسمت رینگ میشه که اونکا رو هم الگوریتم ندیدم pac learnable

حالا باید همه رو ثابت کنیم pac learnable ه.

حالا برای اون قسمت ε بودن رو میاریم و فرمول رو قابل میشیم؟

$$\{S|_x : L_{(D,f)}(h_S) > \varepsilon\} = \{S|_x : S|_x \cap C_1 = \emptyset\}$$

اگر همه داده ها توی رینگ قرار بگیرند باید رینگ با ناحیه C_S اشتراک داشته باشه

اگه اینطوری باشه خطای واقعی مون از ε کمتر میشه

← اینو ادامه میدیم. اون داده‌ها که باهاش اشتراک ندارن نیست $(1-\varepsilon)$ و همین m تا باید

بشه کمتر مساوی $(1-\varepsilon)$

$$D^m(\{S|_x : L_{(D,f)}(h_S) > \varepsilon\}) < (1-\varepsilon)^m \le e^{-\varepsilon m}$$

پس می‌توان گفت H، pac learnable است با:

$$m_H \le \left\lceil \frac{\log(1/\delta)}{\varepsilon} \right\rceil$$

$X=\{0,1\}^d$ $\quad$ $Y=\{0,1\}$ $\quad$ ۳۰۴

یه تابع بولی داریم $f(x)=x_i$ یا $f(x)=1-x_i$

$\wedge$ اگر رو داراپه اگر به اونه به ۱ میبره

یعنی هر x رو میبره به ـــ $x_i \wedge \bar{x}_i$
صفر میبره

میخوایم ثابت کنیم pac learnable است برای اینم در زمینه ثابت کنیم H متناهی است

فرض کنیم $h \in H$ و h حد منفی نباشه.

پس یا x_i رو داره یا $\bar{x}_i$ رو داره یا هیچ کدوم رو نداره

پس $|H| = 3^d + 1$ متناهی است. پس pac learnable شده

$$m_H(\varepsilon,\delta) \leq \left\lceil \frac{\log(|H|/\delta)}{\varepsilon} \right\rceil ,$$

حالا الگوریتم ERM با runtime $d.m$ باشه

فرض کنیم $+ = x_1 \dots x_l$ با استقرا روی l ، h_i رو تعریف کنیم

و میگیم h_i با ERM بله، $\quad h_0(x) := \bigwedge_{j=1}^{d} x_j \bar{x}_j$

$x_{i+1} = (x_1^{i+1}, \dots, x_d^{i+1})$

h_{i+1} رو می‌سازیم

برای هر j ای که اگر $x_j^{i+1} = 1$ ، $\bar{x}_j^{i+1}$ را از h_i حذف می‌کنیم

و اگر برای j ای که $x_j^{i+1} = 0$ آنگاه x_j^{i+1} را از h_i حذف می‌کنیم

حال واضح است h_d به x_1 ، ... ، x_k لیبل مثبت میدهد. چون لیبل مثبت‌ها را

رعایت کردیم. بنابراین این کوچکترین تابع است. پس این الگوریتم ERM است.

Subject: Date: / /

۳۵) فرض کنید X یک دامنه باشد و $D_1, D_2, \ldots, D_m$ یک دنباله از توزیع روی X

$Y = \{0, 1\}$ باشد و فرض کنیم H متناهی هم باشد

و x_1 از D_1 ، x_2 از D_2 ، ...

و S به طول m باشد

$$\bar{D}_m = \frac{D_1 + \cdots + D_m}{m}$$

ثابت کنیم

$\varepsilon \in (0,1)$

$$P\left[\exists h \in H \text{ s.t. } L_{(\bar{D}_m, f)}(h) > \varepsilon \text{ and } L_{(S,f)}(h) = 0\right] \le |H| e^{-\varepsilon m}$$

برای اثبات این نامساوی ابتدا تعریف می‌کنیم H_B آنگاه داریم

$$H_B = \{h \in H : L_{(\bar{D}_m, f)}(h) > \varepsilon\}$$

$$M = \{S|_x : \exists h \in H_B \text{ s.t. } L_{(S,f)}(h) = 0\}$$ و

$$= P(M) = P\left(\bigcup_{h \in H_B} \{S|_x : L_{(S,f)}(h) = 0\}\right)$$

پس $\rightarrow P\left[\exists h \in H \text{ s.t. } L_{(\bar{D}_m, f)}(h) > \varepsilon \text{ and } L_{(S,f)}(h) = 0\right]$

$$\le |H| \times \underbrace{P\left[\{S|_x : L_{(S,f)}(h) = 0\}\right]}$$

حال اگر $h \in H$ وجود داشته باشد که $L_{(\bar{D}_m, f)}(h) > \epsilon$ می‌توان نوشت

$$\frac{P_{x\sim D_1}[h(x) \neq f(x)] + \cdots + P_{x\sim D_m}[h(x) \neq f(x)]}{m} > \epsilon$$

$$\frac{P_{x\sim D_1}[h(x) = f(x)] + \cdots + P_{x\sim D_m}[h(x) = f(x)]}{m} \leq 1 - \epsilon$$

و هم چنین داریم: چون مستقل هستند.

$$P[\{S|_x : L_{(S,f)}(h) = 0\}] = \prod_{i=1}^{m} P_{x\sim D_i}[h(x) = f(x)] =$$

$$\left(\left(\prod_{i=1}^{m} P_{x\sim D_i}[h(x) = f(x)] \right)^{\frac{1}{m}} \right)^{m} \underset{\text{میدانیم}}{\leq} \left(\frac{P_{x\sim D_1}[h(x)=f(x)] + \cdots + P_{x\sim D_m}[h(x)=f(x)]}{m} \right)^{m}$$

$$\leq (1-\epsilon)^m \leq e^{-\epsilon m}$$

پس در آخر می‌توان نوشت

$$P[\exists h \in H \text{ s.t. } L_{(\bar{D}_m, f)}(h) > \epsilon \text{ and } L_{(S,f)}(h) = 0] \leq |H| e^{-\epsilon m}$$

و اثبات کامل است.

۲۰۶ : اگر H، agnostic pac باشد ← H هم pac learnable

هست ؛ و اگر A، الگوریتم agnostic pac باشد آنگاه A pac learnable

هم هست.

فرض کنیم H، agnostic pac باشد، پس الگوریتم A وجود دارد و $m_H:(0,1)^2 \to \mathbb{N}$

هست به طوری که $\epsilon,\delta \in (0,1)$ و توزیع D روی $X \times Y$ باشد:

وطبق تعریف agnostic pac learnable بودن داریم که الگوریتم A، تابع h را برمی‌گرداند که

با احتمال حداقل $1-\delta$ برای $m \geq m_H(\epsilon,\delta)$ ، $L_D(h) \leq \min_{h'\in H} L_D(h') + \epsilon$

حال برای آنکه ثابت کنیم pac learnable هم باید فرض کنیم realizability assumption

رو داریم ، D یک توزیع روی X باشد.

توزیع D' را از D به دست میاریم ، D' روی $X \times \{0,1\}$ با f لیبل می‌زند که

$x \in X$ در D زوج $(x, f(x))$ روی لیبل ۰ . پس $\min_{h'\in H} L_D(h') = 0$

فرض کنیم $\delta, \epsilon \in (0,1)$ با الگوریتم A روی $m \geq m_H(\epsilon,\delta)$ آموزش میدیم

حال با احتمال حداقل $1-\delta$ داریم

$$L_D(h) \leq \underbrace{\min_{h'\in H} L_D(h')}_{0} + \epsilon = \epsilon$$

۵.۲ دو تا الگوریتم داریم ① یه مسئله با دو ویژگی BP و BMI بگیریم
② یه فضای ۵ بعدی که همه چیز و در نظر بگیره بگیریم.

۱. مزایا و معایب هر کدوم از الگوریتم‌ها رو بگو.

در مسئله H_2 داریم و به چیزای محدودی توجه کردیم و ممکنه چیزی رو از دست بدیم و با این کار باعث میشه خطای تقریب بزرگ و خطای تخمین کوچک بشود

و بالعکس وقتی ۵ بعدی رو انتخاب کنیم H بزرگ میشه و ε_{app} کوچیک و ε_{est} بزرگ می‌شود.

۲. توضیح دهید که چگونه تعداد نمونه‌های آموزشی با ترجیح موجود برای انتخاب شما اثر می‌گذارد.

اگر تعداد محدودی از نمونه‌ها داشته باشیم ترجیح میدیم مدلی کوچک‌تر را بگیریم

S کم پس $\downarrow H$ پس $\uparrow$ inductive bias پس $\uparrow \varepsilon_{app}$ و $\downarrow \varepsilon_{est}$

S زیاد پس $\uparrow H$ پس $\downarrow$ inductive bias پس $\downarrow \varepsilon_{app}$ و $\uparrow \varepsilon_{est}$