Subject: Year: Month: Date:

Exercises chapter 3:

1. با توجه به تعریف PAC learnability، به ازای $m \geq m_H(\varepsilon_1,\delta)$
و هر توزیع D روی X، برای یک مجموعه آموزشی S داریم:

$P\{L_D(A(S_m)) < \varepsilon_1\} \geq 1-\delta$

اگر $L_D(A(S_m)) < \varepsilon_1$ باشد، طبق صورت سوال چون $\varepsilon_1 \leq \varepsilon_2$ است
پس $L_D(A(S_m)) < \varepsilon_2$ و طبق تعریف به ازای $m \geq m_H(\varepsilon_2,\delta)$:

$P\{L_D(A(S_m)) < \varepsilon_2\} \geq 1-\delta$

چون $m = m_H(\varepsilon_1,\delta)$ در قسمت اول صدق می‌کند، بنابراین باید در قسمت دوم هم
صدق کند یعنی $m_H(\varepsilon_1,\delta) \geq m_H(\varepsilon_2,\delta)$. برای $0<\delta_1\leq\delta_2<1$
نیز دقیقاً به همین ترتیب است.

2.

1- الگوریتم داده شده برای $H_{singleton}$ به این صورت است، برای هر توزیع D روی X
اگر مجموعه آموزشی S با m عضو داشته باشیم، $S=\{(x_1,f(x_1)),\dots,(x_m,f(x_m))\}$
در صورتی که یک (x_i,y_i) وجود داشته باشد که $y_i=1$، الگوریتم h_{x_i} را برمی‌گرداند
در غیر این صورت الگوریتم h^- را برمی‌گرداند. با توجه به realizability condition

، تنها یک نقطه در X ممکن است وجود داشته باشد که lable داشتن آن 1 باشد

واضح است که الگوریتم برای مجموعه آموزشی همیشه درست عمل می‌کند یعنی $L_S(h_S) = 0$

در نتیجه الگوریتم ERM است.

2- برای اثبات PAC learnability باید ثابت کنیم به ازای $\varepsilon, \delta \in (0,1)$

$$P(\{S_x \mid L_{(D,f)}(h_S) > \varepsilon\}) < \delta$$

که D توزیع دلخواه روی X، f تابع true labling است. با توجه به الگوریتم

اگر $f = h_-$ در این صورت خطای حقیقی همیشه صفر است و شرط بالا برقرار است

هم چنین اگر $f = h_{x_0}$ و $x_0 \in S$، در این حالت باز هم خطای حقیقی

صفر است و شرط بالا برقرار است. بنابراین باید حالتی را بررسی کنیم که $f = h_{x_0}$ اما $x_0 \notin S$.

در این صورت الگوریتم h_- را به عنوان خروجی برمی‌گرداند یعنی $h_S = h_-$. بنابراین

$$P(\{S_x \mid L_{(D,f)}(h_S) > \varepsilon\}) \le P(\{x_0 \notin S_x\})$$

چون اعضای مجموعه S به صورت iid انتخاب می‌شوند، احتمال این که x_0 در S_x

نباشد برابر می‌شود با $1 - P(\{x_0\})$ ← چون m عضو داریم $(1 - P(\{x_0\}))^m$

بنابراین $\varepsilon < L_{(D,f)}(h_-)$

$$L_{(D,f)}(h_-) = P(\{h_-(x) \neq f(x)\}) = P(\{x_0\})$$

یعنی $\varepsilon < P(\{x_0\}) \longrightarrow 1 - P(\{x_0\}) < 1 - \varepsilon$

$\longrightarrow (1 - P(\{x_0\}))^m < (1 - \varepsilon)^m$

بنابراین

$$P(\{S_x \mid L_{(D,f)}(h_-) > \varepsilon\}) \leqslant \underbrace{P(\{x_0 \notin S_x\})}_{=(1-P(\{x_0\}))^m} < (1-\varepsilon)^m < e^{-\varepsilon m}$$

با قرار دادن $e^{-\varepsilon m} \leqslant \delta$ ، و با گرفتن log از طرفین داریم: $m \geqslant \frac{\ln(1/\delta)}{\varepsilon}$

بنابراین برای $m \geqslant \frac{\ln(1/\varepsilon)}{\varepsilon}$ ، خطای حقیقی با احتمال $1-\delta$ از ε کوچکتر است. نقیض گزاره‌ی بالا ثابت شد. $\boxed{\checkmark}$

Subject : Year : Month : Date :

3-

طبق realizability assumption یک h^* وجود دارد که دایره‌ای با $L = 0$

می‌تواند هر نقطه‌ی داخل دایره label 1 ، هر نقطه‌ی خارج از آن label 0 دارد.

شعاع این دایره را r^* می‌نامیم. الگوریتم A را به گونه‌ای در نظر می‌گیریم که کوچک‌ترین دایره‌ای را که تمام نقاط مثبت (۱ دار) مجموعه آموزشی S را در بر می‌گیرد، را باز می‌گرداند. واضح است که

الگوریتم A، ERM است. شعاع این دایره را r_S می‌نامیم. بنابراین $r_S \le r^*$

پس $L_D(h_S) = P_{(x,y)\sim D}(r_S < \|x\| \le r^*)$

برای $\epsilon, \delta \in (0,1)$، فرض می‌دهیم $P(\{x : r_S < \|x\| \le r^*\}) = \epsilon$

و مجموعه‌ی E را به صورت زیر تعریف می‌کنیم.

$E = \{x \in \mathbb{R}^2, r_S \le \|x\| \le r^*\}$

احتمال این که خطای حقیقی بزرگ‌تر از ϵ باشد برابر است با احتمال این که هیچ‌کدام از اعضای

training set عضو مجموعه‌ی E نباشند یعنی $(1-\epsilon)^m$. بنابراین:

$P(\{S_x \mid L_D(h_S) > \epsilon\}) \le P(\{S_x \mid S_x \cap E = \emptyset\})$

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

اگر از طرفین با ساده‌سازی لگاریتم بگیریم داریم $m \ge \frac{\ln(1/\delta)}{\epsilon}$

و ثابت شد با احتمال $1-\delta$ ، خطای حقیقی از ϵ کوچک‌تر است (نقض فرض بالا)

در نتیجه مجموعه فرضیات H ، PAC learnable است.

4-

ابتدا ثابت می‌کنیم که مجموعه فرضیات H متناهی است.

با توجه به صورت سؤال هر h می‌تواند با سه حالت x_i یا $\bar{x}_i$ و یا هیچ‌کدام از این دو

در conjunction مشخص شود. هم چنین تابع All-negative هم وجود دارد

بنابراین اندازه H برابر است با $|H| = 3^d + 1$

چون مجموعه فرضیات H متناهی است بنابراین به ازای $m \geqslant \frac{\log(|H|/\delta)}{\epsilon}$

برای هر توزیع دلخواه D، تابع f، true lable و چون شرط realizability

برقرار است، H، PAC learnable است: $(\epsilon, \delta) \in (0,1)$

$$m \geqslant \frac{\log\left(\frac{3^d+1}{\delta}\right)}{\epsilon} \geqslant \frac{\log(3^d/\delta)}{\epsilon}$$

Sample complexity : $m_H(\epsilon, \delta) \leqslant \frac{\log(3^d/\delta)}{\epsilon} = \frac{d\log 3 + \log(1/\delta)}{\epsilon}$

الگوریتم: h_0 را به صورت زیر تعریف می‌کنیم

$$h_0 = x_1 \wedge \bar{x}_1 \wedge x_2 \wedge \bar{x}_2 \dots \wedge x_d \wedge \bar{x}_d$$

واضح است که تابع h_0 همیشه منفی است. بنابراین اگر y_i در training set

منفی باشد، این تابع همیشه مقدار درست را بر می‌گرداند. برای مقادیر مثبت، باید در تابع

به گونه‌ای تغییر ایجاد کنیم مقادیر مثبت را نیز درست برگرداند تا الگوریتم ERM باشد

Subject: Year: Month: Date:

حال اگر training set به شکل $((a_1, y_1), \dots, (a_d, y_d))$

باشد به ازای $a_i = 1$ ، x_i را حذف می‌کنیم اگر $a_i = 0$ باشد x_i را

حذف نمی‌کنیم و در این صورت اگر a_i مثبت باشد الگوریتم مقدار 1 و در غیر این صورت صفر را

برمی‌گرداند. در نتیجه روی training set به درستی عمل می‌کند و چون شرط realizability

داریم، می‌توانیم نتیجه بگیریم که خطای تجربی h صفر است و الگوریتم ERM است.

Parsian

Subject : Year : Month : Date :

5_

$$L_{(\bar{D}_m, f)}(h) = \frac{1}{m} \sum_{i \in m} L_{(D_i, f)}(h)$$ طبق تعریف صورت سؤال

مجموعه فرضیات بد H_B را در نظر می‌گیریم و داریم: $H_B = \{h \in H \ \ s.t \ \ L_{(\bar{D}_m, f)}(h) > \varepsilon\}$

$$\rightarrow L_{(\bar{D}_m, f)}(h) > \varepsilon \rightarrow \frac{1}{m} \sum_{i \in m} L_{(D_i, f)}(h) > \varepsilon$$

طبق تعریف خطا جمعی:

$$\frac{P_{x \sim D_1}[h(x) = f(x)] + P_{x \sim D_2}[h(x) = f(x)] + \cdots + P_{x \sim D_m}[h(x) = f(x)]}{m} < 1 - \varepsilon \quad *$$

از طرفی
طبق independent بودن
نمونه ها و تعریف خطای کلی

$$P[L_S(h) = 0] = \prod_{i=1}^{m} P_{x \sim D_i}[h(x) = f(x)]$$

Arithmetic_Geometric mean inequality :

$$\frac{a_1 + a_2 + \cdots a_n}{n} \geqslant (a_1 a_2 \ldots a_n)^{1/n}$$

بنابراین ،

$$\prod_{i=1}^{m} P_{x \sim D_i}[h(x) = f(x)] = \left(\left(\prod_{i=1}^{m} P_{x \sim D_i}[h(x) = f(x)]\right)^{1/m}\right)^{m}$$

$$\leqslant \left(\frac{\sum_{i=1}^{m} P_{x \sim D_i}[h(x) = f(x)]}{m}\right)^{m} \underset{*}{\leqslant} (1 - \varepsilon)^{m} \leqslant e^{-\varepsilon m}$$

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

$$P\left(\bigcup_{h \in H} \{L_{(\bar{D}_m, f)}(h) > \varepsilon \text{ and } L_{(S, f)}(h) = 0\}\right)$$

union bound

$$\leqslant P\left(\sum_{h \in H} \{L_{(\bar{D}_m, f)}(h) > \varepsilon \text{ and } L_{(S, f)}(h) = 0\}\right)$$

$$\leqslant \sum_{h \in H} P[L_{(S, f)}(h) = 0] \leqslant |H| e^{-\varepsilon m} \quad \square$$

Parsian

Subject : Year : Month : Date :

6-

اگر مجموعه فرضیات H ، Agnostic PAC learnable باشد یعنی:

$\exists m_H : (0,1)^2 \to \mathbb{N}$ و یک الگوریتم A داشته باشیم به‌طوری‌که برای هر $\varepsilon, \delta \in (0,1)$

و هر توزیع دلخواه D روی $X \times Y$ به‌ازای $m \geq m_H(\varepsilon,\delta)$ نمونه از توزیع D ، با احتمال

$1-\delta$ تابع h را برمی‌گرداند به‌طوری‌که $L_D(h) \leq \min_{h' \in H} L_D(h') + \varepsilon$

اگر علاوه‌بر فرض بالا شرط realizability برقرار باشد، و تابع f. true labling

وجود داشته باشد. بنابراین D یک توزیع دلخواه روی $X \times Y$ است به‌طوری‌که $P_{Y|X}(y|x)$

توسط f مشخص می‌شود. بنابراین D برابر می‌شود با توزیع روی X. هم چنین چون شرط

realizability برقرار است: $\min_{h' \in H} L_{(D,f)}(h') = 0$ بنابراین

با احتمال $1-\delta$: $L_D(h) \leq 0 + \varepsilon \longrightarrow L_D(h) \leq \varepsilon$

7-

: Bayes Factor

$$f_D(x) = \begin{cases} 1 & \text{if } \overbrace{P[y=1|x]}^{\alpha_x} \geq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

$$P[f_D(x) \neq y \mid X=x] = \mathbb{1}_{[\alpha_x \geq 1/2]} \cdot P[Y=0 \mid X=x] + \mathbb{1}_{[\alpha_x < 1/2]} P[Y=1 \mid X=x]$$

$$= \mathbb{1}_{[\alpha_x \geq 1/2]} \cdot (1-\alpha_x) + \mathbb{1}_{[\alpha_x < 1/2]} \alpha_x = \min\{\alpha_x, 1-\alpha_x\}$$

$$\Rightarrow L_D(f_D) = \min\{\alpha_x, 1-\alpha_x\}$$

g را یک تابع classifier دلخواه در نظر می‌گیریم. داریم:

$$L_D(g) = \mathbb{E}_{(x,y)\sim D}\left[\mathbb{1}_{g(x)\neq y}\right] = \mathbb{E}_{x\sim D_x}\left[\mathbb{E}_{y\sim D_{y|x}}\left[\mathbb{1}_{g(x)\neq y}\right]\right]$$

$$= \mathbb{E}_{x\sim D_x}\left[P(g(x)\neq y \mid x)\right]$$

$$P(g(x)\neq y \mid x) = P(g(x)=0 \mid x)\cdot\overbrace{P(y=1\mid x)}^{\alpha_x} + P(g(x)=1\mid x)\,\overbrace{P(y=0\mid x)}^{1-\alpha_x}$$

$$\geqslant P(g(x)=0\mid x)\cdot\min\{\alpha_x, 1-\alpha_x\} + P(g(x)=1\mid x)\cdot\min\{\alpha_x, 1-\alpha_x\}$$

$$= \min\{\alpha_x, 1-\alpha_x\}$$

$$\longrightarrow \mathbb{E}_{x\sim D_x}\left[P(g(x)\neq y\mid x)\right] \geqslant \min\{\alpha_x, 1-\alpha_x\}$$

$$\longrightarrow L_D(g) \geqslant L_D(f_D) \quad \boxed{\checkmark}$$