۲_① اگر S شامل x_+ بود تابع h_{x_+} را برگرداند اگر S شامل x_+ نبود آنگاه تابع h^- را برگرداند. به وضوح الگوریتم پیشنهادی ما ERM است و $L_S(h_S)=0$ است.

۲_② اگر طبق الگوریتم قبلی h^- کار کند خطایی نداریم همچنین داخل فضای S نیز برای x_+ خطایی نداریم حال اگر خارج فضای S، x_+ داشته باشیم فرض کنیم احتمال x_+ با توزیع D، $D(x_+)\leq\varepsilon$ باشد باز خطایی نداریم حال فرض کنیم $D(x_+)\geq\varepsilon$ درنتیجه برای انتخاب نشدن x_+ داریم:

$$D^m(x_+\notin S)\leq(1-\varepsilon)^m\leq e^{-\varepsilon m}$$

پس H_S، PAC است، داریم:

$$e^{-\varepsilon m}\leq\delta \Rightarrow m_{H_S}\leq\left\lceil\frac{\log(1/\delta)}{\varepsilon}\right\rceil$$

۳_② فرض کنیم دایره‌ای با کوچکترین سایز شامل تمام x_+های که هست (ERM) و نام آن را $h^+_{r^+}$ بنامیم از طرفی با فرض assume realizablity می‌دانیم $h^+_{r^*}$ وجود دارد به طوری که خطای آن برابر است با صفر حال چون گفتیم $h^+_{r^+}$ کوچکترین دایره است پس $r^+\leq r^*$ حال تنها جایی که مشکل داریم $r^+<x_+<r^*$ است داریم $D(x_+)\geq\varepsilon$ درنتیجه:

$$D^m(x_+\notin S)\leq(1-\varepsilon)^m<e^{-\varepsilon m}$$ ✓

$$m\leq\left\lceil\frac{\log(\frac{1}{\delta})}{\varepsilon}\right\rceil$$

Scanned with CamScanner

۴- می‌دانیم اگر H متناهی باشد آنگاه $m_H \le \left\lceil \frac{\log |H|/\delta}{\epsilon} \right\rceil$. می‌دانیم H متناهی است و برای هر x_i داریم: $\underbrace{x_i \text{ داخل } h \text{ است}}_{1}$ یا $\underbrace{x_i \text{ داخل } h \text{ است}}_{2}$ یا داخل h نیست

پس اندازه $|H|$ برابر است با 3^d در نتیجه ما نتیجه می‌گیریم H PAC learnable هست و داریم:

$$m_H \le \left\lceil \frac{\log \frac{3^d}{\delta}}{\epsilon} \right\rceil = \left\lceil \frac{d \log \frac{3}{\delta}}{\epsilon} \right\rceil$$

۵- ابتدا h هایی را از H انتخاب می‌کنیم که در آن $L_{(\bar{D}_m, f)}(h) > \epsilon$ که می‌دانیم مقدار کل h هایی که در شرط صدق کند از اندازه‌ی $|H|$ کمتر است. (1)

حال $P[L_{(S,f)}(h)=0]$ را حساب می‌کنیم پس داریم:

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

با استفاده از راهنمایی سوال

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

میانگین توزیع

$$= \left(\sum_{i=1}^{m} P_{x \sim \bar{D}}[h(x)=f(x)] \right)^m \le (1-\epsilon)^m \le e^{-\epsilon m} \quad (2)$$

$$(1),(2) \Rightarrow P[\exists h \in H \text{ s.t. } L_{(\bar{D}_m, f)}(h) > \epsilon \text{ and } L_{(S,f)}(h)=0] \le |H| e^{-\epsilon m}$$

Scanned with CamScanner

۴ـ فرض می‌کنیم H ، agnostic PAC هست و فرض کنیم A هم H را با $m_H(\epsilon,\delta)$ learn می‌کند حال نشان می‌دهیم H با استفاده از A PAC learnable است.

حال توزیع D را روی X فرض کنیم حال فرض کنیم D هست یک توزیع بر روی $X\times\{0,1\}$ که با یک f به طور قطعی تعیین می‌شود در نتیجه ما می‌توان از assume realizability استفاده کنیم و داریم $L_D(h)=0$ ①

حال با استفاده از تعریف agnostic PAC داریم:

$$L_D(h) \le \min_{h'\in H} L_D(h') + \epsilon \xRightarrow[\min L_D(h')=0]{\text{با توجه به ①}}$$

$$L_D(h) \le \epsilon$$

فصل ۵

② من الگوریتم اول که ۲ تا شاخص را در نظر می‌گیرد انتخاب می‌کنم اول از همه پیچیدگی کمتری دارد و در نتیجه خطای کمتری دارد از طرفی ما هرچه متغیرها را بیشتر کنیم H را بزرگ‌تر کرده‌ایم پس در نتیجه امکان overfitting را افزایش داده‌ایم.