تئوری یادگیری ماشین

نیم‌سال اول ۱۴۰۳-۱۴۰۴

مدرس: دکتر امیر نجفی

دانشکده‌ی مهندسی کامپیوتر

تمرین اول

مسئله‌ی ۱. مرور احتمال (۱۰ نمره)

الف) فرض کنید یک تاس منصفانه به شما داده شده است و باید تصمیم بگیرید که کدام وظیفه سخت‌تر است: (i) حدس زدن مقدار یک بار پرتاب تاس یا (ii) دو بار پرتاب تاس و به دست آوردن همان مقدار در هر دو بار. با توجه به اینکه تاس منصفانه است (هر طرف وزن ۱/۶ دارد)، آیا رویداد (i) شانس بیشتری برای موفقیت دارد یا رویداد (ii) یا هر دو احتمال موفقیت یکسانی دارند؟ توضیح دهید.

ب) اکنون این نتیجه را به تاس‌های nـ وجهی با هر توزیع (ممکن است غیر یکنواخت) تعمیم می‌دهیم. ابتدا این حقیقت مفید را ثابت کنید: برای هر $\alpha_1, \alpha_2, \ldots, \alpha_n$ که $\sum_i \alpha_i = 1$، رابطه زیر برقرار است:

$$0 \leqslant \sum_{i=1}^{n} (\alpha_i - 1/n)^2 = \sum_{i=1}^{n} \alpha_i^2 - \frac{1}{n}.$$

ج) بگذارید X_1 مقدار اولین پرتاب و X_2 مقدار دومین پرتاب باشد. نشان دهید که $\Pr(X_1 = X_2) \geqslant 1/n$ (راهنمایی: از قسمت (ب) استفاده کنید). برای چه توزیعی این نابرابری به حداقل می‌رسد؟

حل. الف)پیش‌بینی موفق نتیجه هر یک از رویدادها به طور مساوی محتمل است. به وضوح، احتمال پیش‌بینی مقدار یک پرتاب تاس عادلانه برابر با ۱/۶ است. برای دیدن اینکه احتمال اینکه دو بار مقدار یکسان پرتاب شود چقدر است، فرض کنید X_1 نتیجه اولین پرتاب و X_2 مقدار پرتاب دوم باشد. در این صورت، ما به مقدار زیر علاقه‌مندیم:

$$\Pr(X_1 = X_2) = \sum_{i=1}^{6} \Pr(X_1 = i \wedge X_2 = i).$$

توجه کنید که هر پرتاب مستقل و یکسان است، بنابراین می‌توانیم بنویسیم:

$$\sum_{i=1}^{6} \Pr(X_1 = i \wedge X_2 = i) = \sum_{i=1}^{6} \Pr(X_1 = i) \Pr(X_2 = i) = \sum_{i=1}^{6} \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{6}.$$

ب) این نامساوی درست است زیرا مجموع مربعات همیشه غیرمنفی است. برای نشان دادن تساوی، کافی است عبارات را بسط دهیم و از این واقعیت استفاده کنیم که $\sum_i \alpha_i = 1$. بسط سمت چپ:

$$\sum_{i=1}^{n} (\alpha_i - 1/n)^2 = \sum_{i=1}^{n} \left(\alpha_i^2 - \frac{2}{n}\alpha_i + \frac{1}{n^2} \right) = \sum_{i=1}^{n} \alpha_i^2 - \frac{2}{n} \sum_{i=1}^{n} \alpha_i + \sum_{i=1}^{n} \frac{1}{n^2}.$$

با استفاده از $\sum_{i=1}^n \alpha_i = 1$، این ساده می‌شود به:

$$\sum_{i=1}^n \alpha_i^2 - \frac{2}{n} \cdot 1 + \frac{n}{n^2} = \sum_{i=1}^n \alpha_i^2 - \frac{1}{n}.$$

ج) با تعمیم قسمت (الف) به روشی مستقیم، به رابطه زیر می‌رسیم:

$$\Pr(X_1 = X_2) = \sum_{i=1}^n \Pr(X_1 = i \wedge X_2 = i) = \sum_{i=1}^n \Pr(X_1 = i) \Pr(X_2 = i) = \sum_{i=1}^n \Pr(X_1 = i)^2.$$

توجه کنید که $\sum_{i=1}^n \Pr(X_1 = i) = 1$ طبق تعریف، بنابراین می‌توان $\Pr(X_1 = i) = \alpha_i$ در نظر گرفت. از قسمت (ب) می‌دانیم که $\sum_{i=1}^n \alpha_i^2 \geqslant 1/n$ برقرار است. همانطور که در قسمت (الف) مشاهده شد، توزیع یکنواخت ($\alpha_i = 1/n$) تساوی را محقق می‌کند. ▷

مسئله‌ی ۲. آماره U (۱۰ نمره)

فرض کنید $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ تابعی متقارن و نامنفی از ورودی هایش باشد. با داشتن دنباله‌ای مستقل و هم‌توزیع (i.i.d) از متغیر های تصادفی X_k برای $k \geqslant 1$ عبارت زیر را در نظر بگیرید:

$$U := \frac{1}{\binom{n}{2}} \sum_{j<k} g(X_j, X_k)$$

الف) نشان دهید U تخمین‌گری نااریب برای $\mathbb{E}[g(X_1, X_2)]$ می‌باشد.

ب) فرض کنید برد تابع g همواره کمتر یا مساوی b باشد، نشان دهید:

$$\mathbb{P}[|U - \mathbb{E}[U]| \geqslant t] \leqslant 2e^{-\frac{nt^2}{2b^2}}$$

حل.

الف) با توجه به این که X_i ها مستقل و هم‌توزیع می‌باشند بدیهی است.

ب) ادعا می‌کنیم که U در خاصیت تفاضلات متنهایی صدق می‌کند فرض کنیم X_i را با X_i' عوض کرده‌ایم در این صورت:

$$|U - U'| \leqslant \frac{1}{\binom{n}{2}} \sum_{j \neq i} |g(X_i, X_j) - g(X_i', X_j)| \leqslant \frac{2b}{n}$$

و بنابر نابرابری MacDiarmid حکم نتیجه می‌گردد.

▷

مسئله‌ی ۳. گراف تصادفی (۱۰ نمره)

برای گراف G عدد خوشه‌ای $\mathcal{C}(G)$ را برابر بزرگترین مجموعه از رئوس تعریف می‌کنیم که دوبه‌دو به هم یال داشته باشند. یک گراف تصادفی به این شکل می‌سازیم که بین هر دو راس u, v به طور مستقل از سایر جفت راس ها یک راس به احتمال p قرار می‌دهیم. نشان دهید اگر تعداد یال‌های گراف برابر با n باشد برای این گراف داریم:

$$\mathbb{P}[\frac{1}{n}|\mathcal{C}(G) - \mathbb{E}[\mathcal{C}(G)]| \geqslant \delta] \leqslant 2e^{-2n\delta^2}$$

حل. کافیست توجه کنیم که با تغییر دادن یک یال عدد خوشه‌ای حداکثر به اندازه یک واحد افزایش می‌یابد و بنابر نابرابری MacDiarmid حکم نتیجه می‌گردد. $\triangleright$

مسئله‌ی ۴. تابع غیر قابل یادگیری (۱۰ نمره)

فرض کنید $\mathcal{S}$ کلاس همه‌ی زیرمجموعه‌هایی از $[0, 1]$ باشد که تعداد اعضای آن متناهی است. کلاس توابع $\mathcal{F}$ را به صورت زیر تعریف می‌کنیم:

$$\mathcal{F}_{\mathcal{S}} = \{\mathbb{1}_S(.) | S \in \mathcal{S}\}$$

فرض کنید که X_i ها از توزیعی روی $[0, 1]$ سمپل‌گیری شده‌باشند که تکینگی ندارد. (یعنی $\mathbb{P}(\{x\}) = 0$ برای هر $x \in [0, 1]$) نشان دهید که همگرایی یکنواخت برای خانواده توابع $\mathcal{F}$ و توزیع $\mathbb{P}$ برقرار نیست.

حل. کافیست توجه کنیم که برای هر n سمپل $x_1, \ldots x_n$ یک تابع در خانواده $\mathcal{F}_{\mathcal{S}}$ وجود دارد به طوری که روی $x_1, \ldots x_n$ برابر ۱ باشد در این صورت:

$$\frac{\sum_{i=1}^{n} f(x_i)}{n} = 1, \mathbb{E}f(x) = 0$$

که عبارت دوم به این دلیل است که توزیع ما تکینگی ندارد.

و لذا با افزایش اندازه n مقدار سوپریمم برابر ۱ باقی خواهد ماند و همگرایی یکنواخت نخواهیم داشت. $\triangleright$

مسئله‌ی ۵. دایره‌های هم مرکز (۱۰ نمره)

فرض کنید $\mathcal{X} = \mathbb{R}^2$ و فرض کنید که نقاط کلاس مثبت‌مان توسط دایره‌ای به مرکز مبدا مختصات تعریف شده است. همچنین فرض کنید که فضای فرضیه ما نیز همین دایره های به مرکز مبدا مختصات باشند. نشان دهید این کلاس از مسائل را می‌توان با $\text{PAC-}(\epsilon, \delta)$ با تعداد نمونه $n \geqslant (1/\epsilon)\log(1/\delta)$ یاد گرفت.

حل. کافیست کوچکترین دایره‌ای که شامل همه نقاط است را در نظر بگیریم ادامه راه حل مانند راه حل سوال مستطیل در کلاس است. $\triangleright$

مسئله‌ی ۶. قضیه‌ی گلیونکو-کانتلی (۱۰ نمره)

فرض کنید $X_۱, \ldots, X_n$ نسخه‌هایی مستقل و هم‌توزیع از X باشند که تابع توزیع تجمعی (CDF) آن $F(t) = \mathbb{P}(X \leqslant t)$ است. تابع توزیع تجمعی تجربی X به صورت زیر تعریف می‌شود:

$$\hat{F}_n(t) = \frac{۱}{n}\sum_{i=۱}^{n} \mathbb{I}(X_i \leqslant t)$$

۱. میانگین و واریانس $\hat{F}_n(t)$ را محاسبه کنید و نتیجه بگیرید که $\hat{F}_n(t) \to F(t)$ به صورت تقریباً حتمی وقتی $n \to \infty$. (راهنمایی: از لم بورل‌ـ کانتلی استفاده کنید.)

۲. نشان دهید که برای $n \geqslant ۲$:

$$\mathbb{P}\left(\sup_{t\in\mathbb{R}} \left|\hat{F}_n(t) - F(t)\right| \leqslant C\sqrt{\frac{\log(n/\epsilon)}{n}}\right) = ۱ - \epsilon$$

حل. ۱. ابتدا میانگین و واریانس را محاسبه می‌کنیم:

$$\mathbb{E}[\hat{F}_n(t)] = \frac{۱}{n}\sum_{i=۱}^{n} \mathbb{P}(X_i \leqslant t) = \frac{۱}{n}\sum_{i=۱}^{n} F(t) = F(t),$$

$$\mathrm{Var}[\hat{F}_n(t)] = \frac{۱}{n^۲}\sum_{i=۱}^{n} \mathrm{Var}[\mathbf{۱}(X_i \leqslant t)] = \frac{۱}{n}\mathrm{Var}[\mathbf{۱}(X_۱ \leqslant t)] = \frac{۱}{n}\big(F(t) - F(t)^۲\big).$$

با توجه به اینکه $\hat{F}_n(t)$ میانگین متغیرهای تصادفی مستقل ۰ و ۱ است، می‌توان از نابرابری هافدینگ استفاده کرد: $\epsilon > ۰$

$$\mathbb{P}\big(|\hat{F}_n(t) - F(t)| \geqslant \varepsilon\big) \leqslant ۲\exp(-۲n\varepsilon^۲),$$

$$\sum_{n=۱}^{\infty} \mathbb{P}\big(|\hat{F}_n(t) - F(t)| \geqslant \varepsilon\big) \leqslant \sum_{n=۱}^{\infty} ۲\exp(-۲n\varepsilon) = \frac{۲}{۱ - \exp(-۲\varepsilon)} < \infty,$$

بنابراین، با لم بُرل‌ـ کانتلی، تقریباً با احتمال ۱ تنها تعداد محدودی از وقایع $|\hat{F}_n(t) - F(t)| \geqslant \varepsilon$ رخ می‌دهد. لذا $\hat{F}_n(t) \to F(t)$ تقریباً مطمئن.

۲. می‌توان این را به‌عنوان مسئله‌ای از اندازه تجربی در برابر اندازه واقعی برای کلاس C از بازه‌های نیمه‌باز $(\infty, t]$ با بُعد VC برابر ۱ در نظر گرفت. نابرابری VC اکنون بیان می‌کند که:

$$\sup_{t\in\mathbb{R}} |\hat{F}_n(t) - F(t)| = \sup_{A\in C} |\mu_n(A) - \mu(A)|$$

$$\leqslant ۲\sqrt{\frac{\log(۲en)}{n}} + \sqrt{\frac{\log(۲/\delta)}{n}}$$

$$\leqslant \left(۲\sqrt{\frac{\log(۴e)}{\log ۲}} + \frac{۱}{\sqrt{۲}}\right)\sqrt{\frac{\log(n/\delta)}{n}}$$

برای تمام $n \geqslant ۲$. ▷

مسئله‌ی ۷. یادگیری اتحاد بازه‌ها (۱۰ نمره)

یک الگوریتم یادگیری PAC برای کلاس مفاهیم $C_۲$ که از اجتماع دو بازه بسته به صورت $[a, b] \cup [c, d]$ به طوری که $a, b, c, d \in \mathbb{R}$ تشکیل می‌شود ارائه دهید. نتیجه خود را برای به‌دست‌آوردن یک الگوریتم یادگیری PAC برای کلاس مفهومی C_p گسترش دهید که از اتحاد $p \geqslant ۱$ بازه بسته تشکیل شده است، یعنی $[a_۱, b_۱] \cup \cdots \cup [a_p, b_p]$، با $a_k, b_k \in \mathbb{R}$ برای $k \in [p]$. پیچیدگی زمانی و نمونه‌ای الگوریتم شما به عنوان تابعی از p چیست؟

حل.

راه‌حل: با توجه به یک نمونه S، الگوریتم ما شامل مراحل زیر است:

۱. مرتب‌سازی S به ترتیب صعودی.

۲. گذشتن از داخل S مرتب‌شده و علامت‌گذاری نقاطی که در آن‌ها بازه‌هایی از نقاط مثبت پیوسته شروع و پایان می‌یابند.

۳. ریترن اجتماع بازه‌ها که در مرحله قبل یافت شده است. این اجتماع با یک لیست از دوتایی مرتب نشان داده می‌شود که نقاط شروع و پایان بازه‌ها را مشخص می‌کنند.

این الگوریتم هم برای $p = ۲$ و هم برای p عمومی کار می‌کند. اکنون مسئله را برای $C_۲$ بررسی می‌کنیم. برای نشان دادن اینکه این یک الگوریتم PAC-learning است، نیاز داریم بین دو حالت تمایز قائل شویم.

حالت اول زمانی است که هدف یک اجتماع جدا از دو بازه بسته است: $I = [a, b] \cup [c, d]$. در اینجا دو منبع خطا وجود دارد: خطاهای منفی کاذب در $[a, b]$ و $[c, d]$ و همچنین خطاهای مثبت کاذب در (b, c). خطاهای مثبت کاذب ممکن است زمانی رخ دهند که نمونه‌ای از (b, c) گرفته نشود. با توجه به خطی بودن امید ریاضی و از آنجایی که این دو ناحیه خطا جدا هستند، داریم:

$$R(h_S) = R_{FP}(h_S) + R_{FN,۱}(h_S) + R_{FN,۲}(h_S),$$

که در آن

$$R_{FP}(h_S) = P_{x \sim D}[x \in h_S, x \notin I],$$

$$R_{FN,۱}(h_S) = P_{x \sim D}[x \notin h_S, x \in [a, b]],$$

$$R_{FN,۲}(h_S) = P_{x \sim D}[x \notin h_S, x \in [c, d]].$$

از آنجا که نیاز داریم حداقل یکی از مقادیر $RFP(h_S)$، $RFN_۱(h_S)$، $RFN_۲(h_S)$ از $\epsilon/۳$ بیشتر شود تا $R(h_S) > \epsilon$ باشد، بر اساس کران اجتماع داریم:

$$P(R(h_S) > \epsilon) \leqslant P(RFP(h_S) > \epsilon/۳ \text{ or } RFN_۱(h_S) > \epsilon/۳ \text{ or } RFN_۲(h_S) > \epsilon/۳)$$

$$\leqslant P(RFP(h_S) > \epsilon/۳) + \sum_{i=۱}^{۲} P(RFN_i(h_S) > \epsilon/۳). \qquad (۱)$$

ابتدا $P(RFP(h_S) > \epsilon/۳)$ را محدود می‌کنیم. توجه کنید که اگر $RFP(h_S) > \epsilon/۳$ باشد، آنگاه $P((b,c)) > \epsilon/۳$ و بنابراین:

$$P(RFP(h_S) > \epsilon/۳) \leqslant (۱ - \epsilon/۳)^m \leqslant e^{-m\epsilon/۳}.$$

اکنون می‌توانیم $P(RFN_i(h_S) > \epsilon/۳)$ را با $۲e^{-m\epsilon/۶}$ محدود کنیم، با استفاده از استدلال مشابه در سوال قبلی. بنابراین:

$$P(R(h_S) > \epsilon) \leqslant e^{-m\epsilon/۳} + ۴e^{-m\epsilon/۶} \leqslant ۵e^{-m\epsilon/۶}.$$

قرار دادن طرف راست به صورت δ و حل برای m نشان می‌دهد که $m \geqslant \frac{۶}{\epsilon} \log \frac{۵}{\delta}$.

حالت دومی که باید در نظر بگیریم این است که $I = [a,d]$ باشد، یعنی $[a,b] \cap [c,d] \neq \emptyset$. در این حالت، الگوریتم ما به یادگیری یک بازه $[a,b]$ تقلیل می‌یابد و به سادگی نشان داده می‌شود که تنها $m \geqslant \frac{۲}{\epsilon} \log \frac{۲}{\delta}$ نمونه برای یادگیری این مفهوم کافی است، همانند حالت مستطیل‌های محور_تراز که در کلاس بحث شد. بنابراین، نتیجه می‌گیریم که الگوریتم ما واقعاً یک الگوریتم یادگیری-PAC است.

تعمیم این نتیجه به حالت C_p به صورت مستقیم قابل انجام است. تنها تفاوت این است که در (۱)، دو جمع برای $p - ۱$ ناحیه مثبت کاذب و $۲p$ ناحیه منفی کاذب خواهیم داشت. در این صورت پیچیدگی نمونه $m \geqslant \frac{۲(۲p-۱)}{\epsilon} \log \frac{۳p-۱}{\delta}$ خواهد بود.

گام مرتب‌سازی در الگوریتم ما زمانی $O(m \log m)$ می‌گیرد و دوگام آخر الگوریتم به صورت خطی در m هستند که در مجموع منجر به پیچیدگی زمانی کلی $O(m \log m)$ می‌شود. $\triangleright$

مسئله‌ی ۸. کران‌های تمرکزی (۱۰ نمره)

الف) یک مجموعه از نمونه‌ها به صورت $X = (x_۱, x_۲, \ldots, x_m)$ به طوری که $\forall i, |x_i| \leqslant M$ را در نظر بگیرید. تابع زیر را تعریف کنید:

$$f(X) = \frac{۱}{m} \sum_{i=۱}^{m} x_i.$$

آیا می‌توانید کرانی برای احتمال $\Pr[|f(X) - E[f(X)]| \geqslant \epsilon]$ ارائه دهید؟

ب) فرض کنید X و X' دو مجموعه با اندازه m باشند که دقیقاً در یک نقطه متفاوت‌اند. یعنی $|X \cap X'| = m - ۱$. می‌گوییم یک تابع h پایدار است اگر برای همه‌ی X, X'ها، داشته باشیم $|h(X) - h(X')| \leqslant g(m)$ که g یک تابع نطولی است. سرعت نزول g به عنوان تابعی از m چقدر باید باشد تا نابرابری McDiarmid یک کران برای پیشامد $\Pr[|h(X) - E[h(X)]| \geqslant \epsilon]$ ارائه دهد که با افزایش m به صفر همگرا شود؟

ج) آیا تابع f از قسمت (الف) پایدار است (همچنان فرض می‌کنیم $|x_i| \leqslant M$، $\forall i$)؟ آیا نابرابری McDiarmid یک کران همگرا ارائه می‌دهد؟ اگر بله، کران را بدست آورید. اکنون تابع $f'(X) = \max(X)$ را در نظر بگیرید.

آیا f' پایدار است؟ آیا می‌توانید با نامساوی McDiarmid یک کران ارائه دهید؟

حل. الف)این تنها یک کاربرد مستقیم از نابرابری Hoeffding است. می‌توان تابع f را به‌عنوان مجموع متغیرهای تصادفی x_i/m در نظر گرفت و می‌دانیم که مقدار هر متغیر با $2M/m$ محدود شده است. نابرابری Hoeffding کران زیر را می‌دهد:

$$\Pr[|f(X) - E[f(X)]| > \epsilon] \leqslant 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^m \left(\frac{2M}{m}\right)^2}\right) = 2\exp\left(-\frac{2\epsilon^2 m}{4M^2}\right).$$

ب) برای اینکه نابرابری McDiarmid همگرا شود، نیاز داریم $g(m) \in o(1/\sqrt{m})$. در حالت $g(m) = 1/m^{1/2+\delta}$، می‌توانیم از نابرابری McDiarmid با $c_i = 1/m^{1/2+\delta}$ استفاده کنیم:

$$\Pr[|h(X) - E[h(X)]| \geqslant \epsilon] \leqslant 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^m (m^{-1/2-\delta})^2}\right) = 2\exp\left(-2\epsilon^2 m^{2\delta}\right).$$

به وضوح، نیاز داریم $\delta > 0$ تا کران به صفر همگرا شود زمانی که m به بی‌نهایت میل می‌کند.

ج) تابعی که در قسمت (الف) آمده است پایدار است و $g(m) = 2M/m$ می‌باشد (تغییر یک نقطه محدود باعث تغییر میانگین حداکثر به اندازه $2M/m$ خواهد شد). در واقع، از قسمت (ب) می‌توان نتیجه گرفت که نابرابری McDiarmid، با $c_i = 2M/m$، یک کران همگرا خواهد داد. در حقیقت، در این حالت، کران دقیقاً همان کرانی است که نابرابری Hoeffding ارائه می‌دهد. در اینجا می‌بینیم که نابرابری McDiarmid نابرابری Hoeffding را تعمیم می‌دهد.
تابع f' پایدار نیست و تنها کرانی که می‌توانیم برای آن ارائه دهیم $g(m) \leqslant M$ است (بدون فرضیات دیگر درباره توزیع). بنابراین نمی‌توانیم با استفاده از نابرابری McDiarmid کرانی مفید به دست آوریم. ▷

مسئله‌ی ۹. مدل دو–اوراکل (۱۰ نمره)

در اینجا ما یک سناریوی جایگزین برای یادگیری PAC را در نظر میگیریم که مدل دو اوراکل نامیده می‌شود. فرض کنید که به شما این توانایی داده شده است که به طور صریح درخواست نمونه‌های مثبت یا منفی بدهید، که به ترتیب از توزیع‌های D^+ و D^- کشیده می‌شوند. یک مفهوم به طور کارآمد قابل یادگیری PAC است اگر الگوریتم $\mathcal{A}$ وجود داشته باشد که با داشتن $m = \mathrm{poly}(1/\epsilon, 1/\delta)$ نمونه بتواند یک فرضیه h بدست آورد، که با اطمینان $(1-\delta)$ داشته باشیم $\Pr_{x\sim D^+}[h(x) = 0] \leqslant \epsilon$ و $\Pr_{x\sim D^-}[h(x) = 1] \leqslant \epsilon$. نشان دهید که اگر یک مسئله در معنای کلاسیک به طور کارآمد قابل یادگیر PAC باشد، همیشه در مدل دو اوراکل نیز به طور کارآمد قابل یادگیری PAC است.

حل. فرض کنید c مفهوم حقیقی باشد، در این صورت توجه کنید که:

$$\begin{aligned}
\mathrm{error}(h) &= \Pr_x[h(x) \neq c(x)] \\
&= \Pr_x[h(x) = 1 \wedge c(x) = 0] + \Pr_x[h(x) = 0 \wedge c(x) = 1] \\
&= \Pr_x[h(x) = 1 \mid c(x) = 0]\Pr_x[c(x) = 0] + \Pr_x[h(x) = 0 \mid c(x) = 1]\Pr_x[c(x) = 1].
\end{aligned}$$

فرض کنید $0 < \epsilon < 1/2$، در این صورت طبق فرض می‌دانیم الگوریتمی L وجود دارد که به طور کارآمد فرضیه‌ای h تولید می‌کند به طوری که $\mathrm{error}(h) \leqslant \epsilon/2$ با اطمینان $1 - \delta$. از تساوی‌های بالا نتیجه می‌شود:

$$\Pr_x[h(x) = 1 \mid c(x) = 0] \Pr_x[c(x) = 0] + \Pr_x[h(x) = 0 \mid c(x) = 1] \Pr_x[c(x) = 1] \leqslant \epsilon/2. \quad (1)$$

بنابراین، در مدل دو اوراکلی، می‌توان سناریوی کلاسیک را شبیه‌سازی کرد، به سادگی با در نظر گرفتن نقاط از هر توزیع D^+ یا D^- به عنوان نقاطی که از یک توزیع زیرین مشترک می‌آیند. اگر نقاط را به طور یکنواخت از اوراکل منفی و مثبت بگیریم، خواهیم داشت $\Pr_x[c(x) = 0] = \Pr_x[c(x) = 1] = 1/2$، که به این معنا خواهد بود:

$$\Pr_x[h(x) = 1|c(x) = 0] = \Pr_{x\sim D^-}[h(x) = 1] \leqslant \epsilon,$$

و به طور مشابه $\Pr_x[h(x) = 0 \wedge c(x) = 1] = \Pr_{x\sim D^+}[h(x) = 0] \leqslant \epsilon$. ▷

مسئله‌ی ۱۰. یادگیری PAC برای مستطیل‌های nـ بعدی (۱۰ نمره)

یک الگوریتم یادگیری-PAC برای C، مجموعه مستطیل‌های nـ بعدی هم‌محور در $\mathbb{R}^n$ ارائه دهید، به طوری که $C = \{[a_1, b_1] \times \cdots \times [a_n, b_n] : a_i, b_i \in \mathbb{R}\}$. پیچیدگی نمونه به عنوان تابعی از n چگونه تغییر می‌کند؟ (اثبات شما باید دقیق باشد)

حل. راه‌حل: R' را کوچک‌ترین مستطیل nـ بعدی می‌گیریم که با نمونه داده‌شده سازگار باشد. اثبات مشابه موردی است که در کلاس برای مستطیل‌های در صفحه داده شده است، با این تفاوت که اینجا باید $2n$ ناحیه $r_i, i \in [1, 2n]$ را در امتداد هر وجه مستطیل nـ بعدی هدف در نظر بگیریم، به طوری که $\Pr[r_i] < \frac{\epsilon}{2n}$ و $\Pr[r_i - f_i] < \frac{\epsilon}{2n}$، که در آن f_i وجه داخلی r_i است. همانطور که در اثبات داده شده در کلاس مطرح شد، فرض می‌کنیم $\Pr[R] > \epsilon$، اگر $\Pr[R(R')] > \epsilon$، در این صورت R' باید حداقل یک ناحیه r_i را از دست بدهد. احتمال اینکه هیچ‌یک از m نمونه در ناحیه r_i قرار نگیرد برابر است با $(1 - \epsilon/2n)^m$. با استفاده از کران جمع، نتیجه می‌شود که:

$$\Pr[R(R') > \epsilon] \leqslant 2n(1 - \epsilon/2n)^m \leqslant 2ne^{-\frac{\epsilon m}{2n}}.$$

قرار دادن δ به جای عبارت سمت راست نشان می‌دهد که برای:

$$m \geqslant \frac{2n}{\epsilon} \log \frac{2n}{\delta},$$

با احتمال حداقل $1 - \delta$، $R(R') \leqslant \epsilon$. ▷