فصل هفتم : یادگیری محاسباتی

این فصل به صورت تئوری ویژگی‌های چندین نوع مسئله‌ی یادگیری ماشین و قابلیت‌های چندین نوع از الگوریتم‌های یادگیری ماشین را بیان می‌کند. این تئوری به دنبال جواب سؤالاتی چون "تحت چه شرایطی یادگیری موفق ممکن یا غیرممکن است؟" و "تحت چه شرایطی یک الگوریتم یادگیری خاص موفقیت یادگیری را تضمین می‌کند؟" است. دو چهارچوب[۱] برای بررسی یادگیری الگوریتم‌های یادگیری در نظر گرفته می‌شود. چارچوب اول چهارچوب تقریباً درست[۲] یا (PAC) است، که در آن چارچوب کلاس فرضیه‌هایی را که می‌توان یا نمی‌توان با تعداد چندجمله‌ای‌ای از نمونه‌های آموزشی یاد گرفت را بررسی و معیاری طبیعی برای پیچیدگی فضای فرضیه‌ای که تعداد نمونه‌های آموزشی برای یادگیری استقرایی را محدود می‌کند تعریف خواهیم کرد. در چارچوب کران خطا[۳] تعداد خطاهای آموزشی‌ای را که یادگیر قبل از تعیین فرضیه‌ی درست انجام می‌دهد را بررسی خواهیم کرد.

۷,۱ مقدمه

در مطالعه‌ی یادگیری ماشین این سؤال طبیعی است که بپرسیم چه قوانین کلی‌ای بر یادگیرهای ماشین (یا غیر ماشین) حاکم است. آیا می‌توان کلاس‌های مسائل یادگیری را که ذاتاً سخت یا آسان‌اند را مستقل از الگوریتم یادگیری تعیین کرد؟ آیا می‌توان تعداد نمونه‌های لازم برای اینکه یادگیری حتماً موفق باشد را تعیین کرد؟ اگر یادگیر بتواند بجای آموزش با دسته‌ی معینی از نمونه‌ها آزمایش انجام دهد (در مقابل اینکه نمونه‌ها به صورت تصادفی به یادگیر داده شوند) این تعداد چگونه تغییر خواهد کرد؟ یا آیا می‌توان تعداد خطاهای یادگیر قبل از یادگیری تابع هدف را مشخص کرد؟ آیا می‌توان پیچیدگی محاسباتی ذاتی کلاس‌های مسائل مختلف را مشخص کرد؟

---

[۱] framework
[۲] probably approximately correct
[۳] mistake bound framework

اگر چه جواب جامع همه‌ی این سؤالات هنوز معلوم نیست، اما قسمت از تئوری هوش محاسباتی برای پاسخ به این سؤالات به وجود آمده است. این فصل نتایج کلیدی این تئوری و جواب به این سؤالات در بعضی مسائل خاص را در بر می‌گیرد. در اینجا بحث را به مسئله‌ی یادگیری استقرایی تابع هدفی نامعلوم از نمونه‌های آموزشی این تابع هدف و فضای فرضیه‌ای معلوم محدود می‌کنیم. با این تعریف مسئله، پاسخ به سؤالاتی مثل تعداد نمونه‌های لازم برای یادگیری موفق و تعداد اشتباهات قبل از یادگیری کامل مطرح می‌شود. همان‌طور که بعداً نیز خواهیم دید تعیین مرزهای این کمیت‌ها به ویژگی‌های مسئله‌ی یادگیری از جمله موارد زیر وابسته است:

- اندازه یا پیچیدگی فضای فرضیه‌ای در نظر گرفته شده
- دقت لازم برای یادگیری
- احتمال اینکه یادگیر فرضیه‌ای موفق را خروجی دهد
- روند ارائه‌ی نمونه‌ها

در اکثر موارد، ما بر روی الگوریتم یادگیری خاصی تمرکز نمی‌کنیم و ترجیح داده می‌شود بیشتر بر روی کلاس‌های الگوریتم‌های یادگیری با خواص یکسان (فضای فرضیه‌ای مشابه، نحوه‌ی نمایش نمونه‌های آموزشی مشابه و ...) بحث شود. هدف از این فصل پاسخ به سؤالاتی نظیر سؤالات زیر است:

- پیچیدگی نمونه‌ای[۱]. تعداد نمونه‌های آموزشی لازم برای اینکه یادگیر (با احتمال بالایی) به فرضیه‌ای موفق میل کند؟
- پیچیدگی محاسباتی[۲]. چه میزان محاسبه انجام می‌شود تا یادگیر با احتمال خوبی به فرضیه‌ای موفق میل کند؟
- مرز خطا[۳]. تعداد نمونه‌ها آموزشی‌ای که یادگیر قبل از همگرا شدن به فرضیه موفق غلت دسته‌بندی می‌کند؟

توجه داشته باشید که در بسیاری از حالات چنین سؤالاتی مطرح‌اند. برای مثال، روش‌های گوناگونی برای تعریف "موفق" وجود دارد. ممکن است یادگیری فرضیه‌ای را موفق تعریف کنیم که فرضیه‌ی خروجی‌اش دقیقاً مشابه مفهوم هدف باشد. یا در مقابل ممکن است یادگیری را موفق بدانیم که فرضیه‌اش در اکثر مواقع مشابه مفهوم هدف باشد، یا به طور معمول چنین فرضیه‌ای را خروجی می‌دهد. یا به طور مشابه، روند ارائه‌ی نمونه‌ها ممکن است متفاوت باشد، ممکن است این نمونه‌ها توسط یک معلم به یادگیر داده شود یا یادگیر اجازه‌ی انجام آزمایش داشته باشد یا اینکه نمونه‌ها توسط یک فرایند تصادفی خارج از کنترل یادگیر انتخاب شوند. همان‌طور که انتظار می‌رود، جواب این سؤالات به تعریف مسئله و مدل یادگیری وابسته است.

ادامه‌ی این فصل به صورت زیر ساختاربندی شده است. قسمت ۷٫۲ حالت یادگیری احتمالی تقریباً درست (PAC) را معرفی می‌کند. در ادامه، قسمت ۷٫۳ پیچیدگی نمونه‌ای و پیچیدگی محاسباتی چندین مسئله‌ی یادگیری را در این حالت بررسی می‌کند. قسمت ۷٫۴ معیار مهمی از پیچیدگی فضا[۴] به نام بعد VC و تأثیر آن در بررسی PACمان در مسائلی که فضای فرضیه محدود است را بررسی خواهیم کرد. قسمت ۷٫۵ مدل مرز خطا را معرفی کرده و مرزی برای تعداد خطاهای الگوریتم‌های مختلف یادگیری فصول قبلی پیدا می‌کند. در انتها نیز، الگوریتم Weighted-Majority را معرفی می‌کنیم، این الگوریتم روشی برای تلفیق پیش‌بینی‌های الگوریتم‌های مختلف رقیب است، مرز خطای تئوری این الگوریتم را نیز بررسی خواهیم کرد.

---

۱ Sample complexity
۲ Computational complexity
۳ Mistake bound
۴ space complexity

۷,۲ احتمال یادگیری یک فرضیه‌ی تقریباً درست

در این بخش حالت خاصی را برای مسائل یادگیری در نظر می‌گیریم، این حالت مدل یادگیری تقریباً درست (PAC) نامیده می‌شود. بیایید کار را با این سؤال که چه تعداد نمونه‌ی آموزشی و چه میزان محاسبه لازم است تا کلاس‌های مختلف یادگیری را با این مدل یاد بگیریم شروع کنیم. برای سادگی کار، بحث را به یادگیری مفاهیم منطقی از داده‌های آموزشی بدون خطا محدود می‌کنیم. با این وجود بسیاری از نتایج حاصل را می‌توان به حالت کلی یادگیری توابع حقیقی مقادیر تابع هدف تعمیم داد (برای مثال به (Natarajan 1991) مراجعه کنید) و بسیاری دیگر از نتایج را می‌توان به یادگیری از انواع خاصی از داده‌های خطادار تعمیم داد (برای مثال بـه (Laird 1988) و (Kearns and Vazirani 1994) مراجعه کنید).

۷,۲,۱ تعریف مسئله

مشابه فصول گذشته، X مجموعه‌ی تمامی نمونه‌های ممکن بر روی تابع هدف مفروض است. برای مثال، X ممکن است مجموعه‌ی تمامی افراد باشد که با ویژگی‌های (age (young or old و (height (short or tall باشد. C مجموعه‌ی مفاهیم هدفی است که ممکن است یادگیر برای یادگیری آن‌ها به کار برده شود. هر مفهوم هدف c در C متناسب با زیرمجموعه‌ای از X است یا به طور مشابه متناسب با تابع $c: X \rightarrow \{0,1\}$ است. برای مثال، یک تابع هدف c در C ممکن است مفهوم "افراد اسکی‌باز" باشد. اگر x نمونه‌ی مثبتی از c باشد، داریم که $c(x)=1$؛ و اگر x نمونه‌ی منفی‌ای باشد داریم $c(x)=0$.

در این حالت فرض می‌کنیم که نمونه‌ها به صورت تصادفی و با توزیع احتمال $\mathcal{D}$ انتخاب می‌شوند. برای مثال، $\mathcal{D}$ ممکن است توزیع احتمال نمونه‌ها افرادی باشد که از یک باشگاه ورزشی در سوئد بیرون می‌آیند (توزیع احتمالی بر روی تمامی افراد). در کل $\mathcal{D}$ ممکن است هر توزیع احتمالی باشد و در حالت کلی این توزیع احتمال برای یادگیر ناشناخته است. تمامی اطلاعات موجود در مورد $\mathcal{D}$ این است که توزیع احتمالی ثابت است؛ بدین معنا که این توزیع احتمال با زمان تغییر نمی‌کند. نمونه‌های آموزشی با این توزیع احتمال انتخاب شده و به همراه مقدار تابع هدفشان $c(x)$ به یادگیر داده می‌شوند.

یادگیر L مجموعه‌ای از فرضیه‌های ممکن مثل H را در یادگیری مفهوم هدف در نظر می‌گیرد. برای مثال، H ممکن است مجموعه‌ی تمامی فرضیه‌های قابل‌بیان به صورت عطف ویژگی‌های age و height باشد. بعد از مشاهده‌ی سری‌ای از نمونه‌های آموزشی برای تابع هدف c، L باید فرضیه‌ای مثل h از H که تخمین آن از c است به عنوان فرضیه‌ی تخمینی خروجی دهد. موفقیت L را کارایی این فرضیه h بر روی نمونه‌های جدیدی که به صورت تصادفی از X و با توزیع $\mathcal{D}$ انتخاب می‌شوند می‌سنجیم. توزیع احتمال $\mathcal{D}$ همان توزیع احتمالی است که نمونه‌های آموزشی با انتخاب شده‌اند.

در چنین حالتی، علاقه‌ی ما به بررسی کارایی یادگیرهای مختلف L با فضای فرضیه‌های مختلف H در یادگیری مجموعه توابع هدف مختلف درون C است. زیرا که می‌خواهیم یادگیر L به اندازه‌ی کلی جامع باشد تا بتواند هر تابع هدف درون C را مستقل از اینکه توزیع $\mathcal{D}$ چیست یاد بگیرد. در بعضی مواقع نیز علاقه داریم که در بدترین حالت توابع هدف درون C را برای تمامی توزیع‌های $\mathcal{D}$ را بررسی کنیم.

۷٫۲٫۲ خطای یک فرضیه

چون علاقه‌ی ما به نزدیکی فرضیه خروجی یادگیر h به تابع هدف حقیقی c است، بیایید کار را با تعریف خطای واقعی[۱] یک فرضیه‌ی h بر روی c و توزیع احتمال $\mathcal{D}$ شروع کنیم. به صورت غیررسمی خطای واقعی h، خطای h در دسته‌بندی نمونه‌های جدید با توزیع $\mathcal{D}$ است. در واقع این تعریف خطا همان تعریف خطای فصل ۵ است. برای راحتی تعریف را برای c که مفهومی منطقی است بازنویسی می‌کنیم.

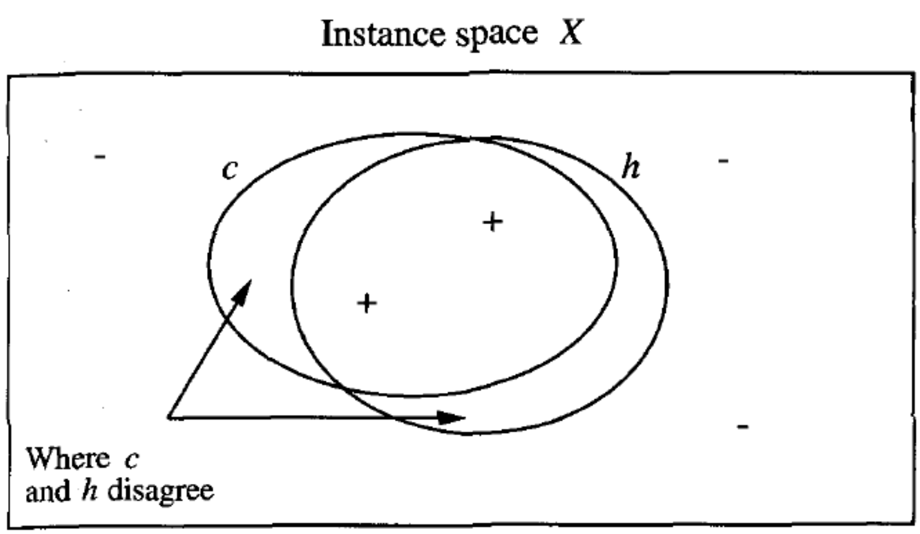
تعریف: خطای واقعی ($error_D(h)$)ی فرضیه‌ی h برای تابع هدف c و توزیع احتمال نمونه‌ای $\mathcal{D}$ احتمال این است که نمونه‌ی انتخابی بر اساس توزیع $\mathcal{D}$ اشتباه دسته‌بندی شود.

$$error_D(h) \equiv \Pr_{x \in \mathcal{D}} [c(x) \neq h(x)]$$

در اینجا نماد $\Pr_{x \in D}$ نشان‌دهنده‌ی احتمال عبارت با فرض اینکه x از توزیع $\mathcal{D}$ پیروی می‌کند است.

شکل ۷٫۱ این تعریف را به فرم گرافیکی نشان می‌دهد. مفاهیم c و h با مجموعه‌ی نمونه‌های X نمایش داده شده‌اند، نمونه‌های آموزشی در این مثال با علامت‌های + و − نشان داده شده‌اند. خطای h برای c احتمال دسته‌بندی غلت نمونه تصادفی در این صفحه یا قرار گرفتن در اختلاف این دو مجموعه (قرار گرفتن در ناحیه هلالی) است. توجه دارید که خطا را طوری تعریف کرده‌ایم که خطای تمامی نمونه‌های ممکن را اندازه بگیرد و فقط محدود به نمونه‌های آموزشی نباشد بنابراین انتظار داریم که زمانی که از فرضیه به دست آمده بر روی نمونه‌های تصادفی جدید استفاده می‌کنیم چنین خطایی داشته باشند.

توجه دارید که این خطا به شدت به توزیع احتمال نامعلوم $\mathcal{D}$ وابسته است. برای مثال اگر $\mathcal{D}$ توزیعی یکنواخت باشد که به تمامی نمونه‌های X احتمالی یکسان نسبت می‌دهد خطای فرضیه‌ی آمده در شکل ۷٫۱ نسبت نمونه‌های درون ناحیه هلالی به تمامی نمونه‌ها خواهد بود. با این وجود اگر $\mathcal{D}$ احتمال بیشتری به نمونه‌های ناحیه هلالی نسبت دهد این خطا بیشتر خواهد شد. و در بدترین حالت $\mathcal{D}$ احتمال صفر به نمونه‌های خارج ناحیه هلالی نسبت می‌دهد و خطا ۱ خواهد بود با وجود اینکه h و c واقعاً اشتراک دارند.



[۱] True Error

شکل ۷٫۱ خطای فرضیه h برای مفهوم هدف c.
خطای h برای مفهوم هدف c احتمال این است که نمونه‌ای تصادفی در درون ناحیه‌ای قرار بگیرد که h و c در آن ناحیه دسته‌بندی‌های مشابهی ندارند. نقاط + و – نشان‌دهنده‌ی نمونه‌های آموزشی مثبت و منفی‌اند. توجه داشته باشید که با وجود اینکه در تمامی ۵ نمونه‌ی مشاهده شده دسته‌بندی h و c یکی است، h خطای غیر صفری برای مفهوم هدف c دارد.

بالاخره، توجه داشته باشید که این خطای h برای c به طور مستقیم برای یادگیر غیرقابل مشاهده است. L فقط کارایی h بر روی نمونه‌های آموزشی را در دسترس دارد و باید انتخاب خود در مورد فرضیه را بر اساس همین معیار انجام دهد. ما از عبارت خطای آموزشی[۲] (در مقابل خطای واقعی) برای نمایش نسبت نمونه‌های آموزشی با دسته‌بندی اشتباه توسط h به کل نمونه‌های آموزشی استفاده می‌کنیم. قسمت بزرگی از بررسی ما از پیچیدگی یادگیری بر محور این سؤال متمرکز می‌شود که "چگونه احتمال دارد که خطای آموزشی مشاهده شده معیاری غلت انداز از $error_{\mathcal{D}}(h)$ باشد؟" است.

به رابطه‌ی بین این سؤال و سؤال مطرح شده در فصل ۵ دقت کنید. با توجه به آنچه در فصل ۵ گفته شد، خطای نمونه‌ای h را برای مجموعه‌ی S از نمونه‌ها نسبت دسته‌بندی اشتباه اعضای S توسط h تعریف شد. خطای آموزشی تعریف شده در بالا خطای نمونه‌ای S است با این فرض که S مجموعه‌ی نمونه‌های آموزشی باشد. در فصل ۵ احتمال اینکه خطای نمونه‌ای تخمینی غلت انداز از خطای واقعی باشد را با این فرض که داده‌های نمونه‌ی S مستقل از h باشند بررسی کردیم. اما اینجا حتی این فرض هم درست نیست و فرضیه‌ی h کاملاً وابسته به مجموعه S است! بنابراین، در این فصل ما این حالت خاص مهم را بررسی خواهیم کرد.

۷٫۲٫۳ قابلیت یادگیری PAC

هدف ما تعیین ویژگی‌های توابع هدفی است که می‌توان آن‌ها را از تعداد معقولی نمونه آموزشی تصادفی با پیچیدگی محاسباتی معقولی یاد گرفت.

چه عبارت‌هایی را می‌توان درباره‌ی قابلیت یادگیری یک تابع بیان کرد درست فرض کرد؟ ممکن است سعی کنیم تعداد نمونه‌های آموزشی لازم برای یادگیری فرضیه‌ای با $error_{\mathcal{D}}(h) = 0$ را تعیین کنیم. متأسفانه در این تعریف مسئله به دو دلیل این کاری بیهوده است. ابتدا اینکه برای اینکه به چنین خطایی برسیم باید تمامی نمونه‌های X را به عنوان نمونه‌ی آموزش به یادگیر ارائه کنیم (که این فرضی غیرواقعی است)، و ممکن است چندین فرضیه با مجموعه نمونه‌های آموزشی سازگار باشند و یادگیر در انتخاب فرضیه‌ی تخمینی برای مفهوم هدف سردرگم خواهد ماند. دوم اینکه با معلوم بودن نمونه‌های آموزشی تصادفی، همیشه احتمالی غیر صفر وجود دارد که نمونه‌های آموزشی معیاری غلت انداز باشد. (برای مثال، با وجود اینکه اغلب قد افراد خارج شده از یک مجموعه‌ی ورزشی در سوئد متفاوت است اما احتمال غیر صفری وجود دارد که در یک روز تمامی نمونه‌های مشاهده شده قد ۲ متر داشته باشند).

برای غلبه بر این دو مشکل، شرایط خواستاری مسئله را از دو نظر کاهش می‌دهیم. ابتدا بجای اینکه شرط کنیم خطای فرضیه صفر شود شرط می‌کنیم که خطا از مقدار دلخواه کوچک ε کوچک‌تر باشد. دوم اینکه بجای اینکه شرط کنیم یادگیر روی هر نمونه‌ی آموزشی ممکن موفق باشد شرط می‌کنیم که احتمال عدم موفقیت کمتر از حد دلخواه کوچک خاصی، δ، کمتر باشد. به طور خلاصه شرط می‌کنیم که یادگیر به صورت احتمالی فرضیه‌ای تقریباً درست[۳] را یاد بگیرد، بنابراین به این فرضیه، فرضیه‌ی احتمالی تقریباً درست یا PAC می‌گویند.

---

[۲] training error
[۳] approximately correct

مجموعه‌ی C را به عنوان مجموعه‌ی مفاهیم هدف ممکن و یادگیر L با فضای فرضیه‌ای H را در نظر بگیرید. به طور غیررسمی، زمانی می‌گوییم که مجموعه‌ی C توسط L با استفاده از H قابل یادگیری PAC [۴] است که، برای هر تابع هدف c در C، L با احتمال $(1-\delta)$ با داشتن تعداد قابل قبولی نمونه‌ی آموزشی و انجام مقدار قابل قبولی محاسبه فرضیه‌ای مثل h خروجی دهد که داشته باشیم، $error_{\mathcal{D}}(h) < \varepsilon$. به عبارت دقیقتر،

تعریف: مجموعه‌ی مفاهیم هدف C را که بر روی نمونه‌های X با اندازه‌ی n تعریف شده و یادگیر L با فضای فرضیه‌ای H را در نظر بگیرید. C توسط L با استفاده از H زمانی قابل یادگیری PAC است که برای تمامی $c \in C$ و توزیع‌های $\mathcal{D}$ بر روی X و $0 < \varepsilon < 1/2$ و δ هایی که $0 < \delta < 1/2$، یادگیر L با احتمال حداقل $(1-\delta)$ فرضیه $h \in H$ خروجی دهد که داشته باشیم که $error_{\mathcal{D}}(h) \le \varepsilon$، در زمان حداکثر چندجمله‌ای از $1/\delta$، $1/\varepsilon$، n و size(c) خروجی باید محاسبه شود.

تعریف ما دو شرط بر روی L می‌گذارد. ابتدا اینکه L باید با احتمال دلخواه بالایی $(1-\delta)$ فرضیه‌ای با مقدار خطای به اندازه‌ی دلخواه کوچکی ε خروجی دهد. دوم اینکه باید کارایی خوبی داشته باشد، در زمانی حداکثر چندجمله‌ای از $1/\delta$، $1/\varepsilon$، n و size(c) فرضیه‌ی خروجی را مشخص کند. در اینجا، n تعداد نمونه‌های X است. برای مثال، اگر نمونه‌های X، عطف k ویژگی منطقی باشند، خواهیم داشت که n=k. پارامتر دوم فضا، size(c) است که اندازه‌ی کدهای c های درون C را نشان می‌دهد با فرض اینکه برای C نمایش خاصی را تعیین کرده باشیم. برای مثال اگر مفاهیم C با عطف حداکثر k ویژگی منطقی باشند، که هر کدام با لیستی از ویژگی‌ها مشخص شود، size(c) تعداد ویژگی‌های واقعی به کار رفته در توضیح c خواهد بود.

ممکن است ابتدا به نظر برسد که تعریف ما از یادگیری PAC فقط اهمیت منابع محاسباتی لازم برای یادگیری در نظر گرفته شده است در حالی که در عمل تعداد نمونه‌های لازم برای یادگیری بیشتر برای ما اهمیت دارد. با این وجود، این دو بسیار به یکدیگر نزدیک‌اند: اگر L نیاز به پردازش حداقلی برای هر نمونه‌ی آموزشی داشته باشد، برای اینکه C را قابل یادگیری PAC توسط L بدانیم، L حتماً به تعداد چندجمله‌ای‌ای نمونه‌ی آموزشی نیاز خواهد داشت. در اصل روشی متداول برای نشان دادن اینکه C ای خاص قابل یادگیری PAC است این است که ابتدا نشان دهیم هر مفهوم هدف از تعداد چندجمله‌ای‌ای از نمونه‌های آموزشی قابل یادگیری است و سپس نشان دهیم زمان محاسبات نیز از چندجمله‌ای کمتر است.

قبل از رفتن به قسمت بعد، باید به فرض محدود کننده‌ای در تعریفمان از قابل یادگیری PAC اشاره کنیم. این تعریف مطلقاً فرض می‌کند که فضای فرضیه‌ای یادگیر H شامل فرضیه‌هایی است که خطای به اندازه‌ی دلخواه کوچک برای تمامی مفاهیم درون C دارند. این فرض از این حقیقت ناشی می‌شود که در تعریف بالا یادگیر زمانی موفق است که بتوان ε را به اندازه‌ی دلخواه به صفر نزدیک کرد. البته در حالتی که C دقیق معلوم نیست (برای مثال، C در برنامه‌ای که باید تصاویر چهره را تشخیص دهد چیست؟) تضمین این شرط سخت خواهد بود، مگر اینکه H مجموعه‌ی توانی X در نظر گرفته شود. همان‌طور که در فصل ۲ نیز گفته شد، چنین H بدون بایاسی دقت کافی تعمیمی با تعداد قابل قبولی از نمونه‌های آموزشی پیدا نمی‌کند. با این وجود، نتایج حاصل از مدل یادگیری PAC دید مفیدی درباره‌ی پیچیدگی نسبی مسائل یادگیری مختلف و ضریب بهبود تعمیم با افزایش نمونه‌های آموزشی به ما می‌دهد. علاوه بر این، در بخش ۷،۳،۱ این فرض محدود کننده را برای در نظر گرفتن یادگیر بدون پیش‌فرض حذف می‌کنیم.

---

[۴] PAC learnable

۷,۳ پیچیدگی نمونه‌ای برای فضای فرضیه‌ای محدود

همان‌طور که در بالا نشان دادیم، قابلیت یادگیری PAC [۵] به شدت به تعداد نمونه‌های آموزشی لازم برای یادگیر وابسته است. افزایش تعداد نمونه‌های آموزشی لازم برای یادگیری متناسب با اندازه‌ی مسئله، که پیچیدگی نمونه‌ای مسئله‌ی یادگیری نامیده می‌شود، از مهم‌ترین معیارهای مسائل یادگیری است. علت این اهمیت از این رو است که در مسائل عملی محدودیت موفقیت در یادگیری بیشتر به خاطر محدودیت یادگیر در تعداد نمونه‌های آموزشی است.

در اینجا ما مرزی کلی برای پیچیدگی نمونه‌ای برای کلاس بزرگی از یادگیرها، یادگیرهای سازگار[۶] ارائه می‌کنیم. یادگیر سازگار یادگیری است که زمانی که ممکن باشد فرضیه‌ای را که با نمونه‌های آموزشی سازگار باشد خروجی می‌دهد. انتظار سازگار بودن یادگیرها انتظاری دور از ذهن نیست، زیرا که معمولاً ما فرضیه‌ای را که با نمونه‌های آموزشی سازگار است را به فرضیه‌های دیگر ترجیح می‌دهیم. توجه دارید که اکثر الگوریتم‌های مطرح شده در فصول قبلی از جمله تمامی الگوریتم‌های فصل ۲ سازگار هستند.

آیا می‌توانیم مرزی برای تعداد نمونه‌های آموزشی لازم برای هر یادگیر سازگار که مستقل از الگوریتم است پیدا کنیم؟ جواب این سؤال بلی است. برای ایجاد چنین مرزی بد نیست که بعضی تعاریف فصل ۲ را درباره‌ی فضای ویژه بازگو کنیم. در آنجا فضای ویژه‌ی $VS_{H,D}$ را مجموعه‌ی تمامی فرضیه‌های $h \in H$ را که نمونه‌های آموزشی D را درست دسته‌بندی می‌کنند تعریف کردیم.

$$VS_{H,D} = \{\mathrm{h} \in H | (\forall < x, c(x) > \in D)(\mathrm{h(x)} = \mathrm{c(x)})\}$$

اهمیت فضای ویژه در اینجا این است که هر یادگیر سازگار مستقل از اینکه X یا H یا D چه باشند فرضیه‌ای از فضای ویژه را خروجی می‌دهد. دلیل این نتیجه در تعریف فضای ویژه مشهود است، زیرا که فضای ویژه تمامی فرضیه‌های سازگار در H با نمونه‌های آموزشی را در بر می‌گیرد. بنابراین برای محدود کردن تعداد نمونه‌های آموزش لازم برای هر یادگیر سازگار کافی است تعداد نمونه‌های آموزشی لازم را برای تضمین اینکه فضای ویژه فرضیه‌ای غیرقابل‌قبول را در بر نگیرد معلوم کنیم. تعریف زیر، که به نام Haussler (1988) نام‌گذاری شده، این شرط را به صورت دقیق مشخص می‌کند.

تعریف: فضای فرضیه‌ای H، مفهوم هدف c، توزیع نمونه‌ای $\mathcal{D}$ و مجموعه‌ی نمونه‌های آموزشی D که برای آموزشی c است را در نظر بگیرید. فضای ویژه‌ی $VS_{H,D}$ زمانی ε-exhausted برای c و D است که برای هر فرضیه‌ی h در $VS_{H,D}$ خطایی کمتر از ε برای c و D داشته باشیم.

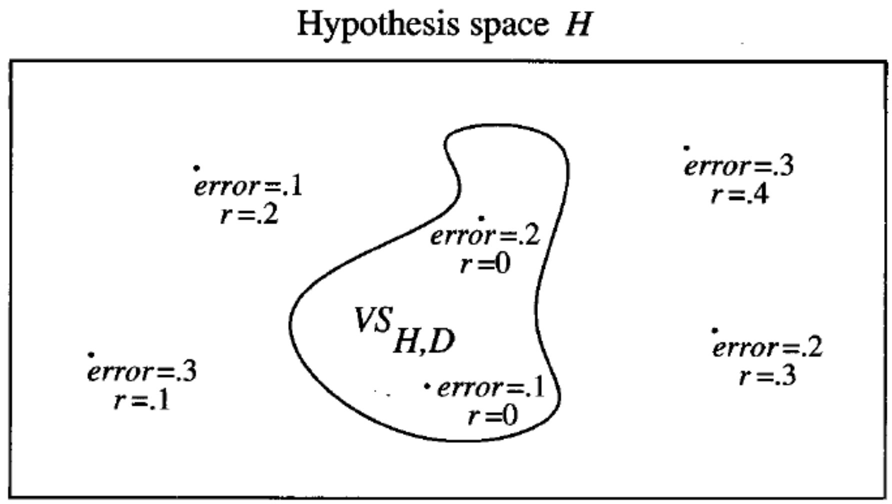
$$\left(\forall h \in VS_{H,D}\right) error_{\mathcal{D}}(h) < \varepsilon$$

تعریف بالا در شکل ۷,۲ نمایش داده شده است. فضای ویژه زمانی ε-exhausted است که تمامی فرضیه‌های سازگار با نمونه‌های آموزشی مشاهده شده (برای مثال، آن‌هایی که خطای نمونه‌ای صفر دارند) خطایی کمتر از ε داشته باشند. البته از دید یادگیر فقط فرضیه‌هایی که به طور کامل با نمونه‌های آموزشی سازگارند قابل‌تشخیص است، همگی آن‌ها خطای آموزشی صفر خواهند داشت. فقط شاهدی که از ماهیت مفهوم هدف آگاه است می‌تواند با قطعیت فضای ویژه‌ی ε-exhausted را مشخص کند. جالب است که بررسی‌ای احتمالی به ما اجازه می‌دهد که

[۵] PAC-learnability

[۶] consistent learners

احتمال اینکه فضای ویژه بعد از تعدادی نمونه‌ی آموزشی ε-exhausted باشد را بدون اینکه اطلاعاتی در مورد ماهیت مفهوم هدف یا توزیع نمونه‌های آموزشی داشته باشیم محدود کنیم. Haussler (1988) چنین مرزی را با قضیه‌ی زیر ایجاد می‌کند.



شکل ۷,۲ exausted کردن فضای ویژه.
فضای ویژه‌ی $\boldsymbol{VS_{H,D}}$ زیرمجموعه‌ای از فرضیه‌های h∈H است که خطای آموزشی صفر دارند (در شکل با r=0 نشان داده شده است). البته خطای واقعی $\boldsymbol{error_D(h)}$ است (که در شکل با error نمایش داده شده) که حتی ممکن است برای فرضیه‌های فضای ویژه غیر صفر باشد. فضای ویژه زمانی ε-exausted است که تمامی فرضیه‌های باقیمانده‌ی درون $\boldsymbol{VS_{H,D}}$ داشته باشیم $\boldsymbol{error_D(h) < \varepsilon}$.

قضیه‌ی ۷,۱ فضای ویژه‌ی ε-exhausted. اگر فضای فرضیه‌ای H محدود باشد و D نیز سرای ای از m≥1 نمونه‌ی تصادفی مستقل از مفهوم هدف c باشد، برای هر 0≤ε≤1 احتمال اینکه فضای ویژه‌ی $VS_{H,D}$ برای c ε-exhausted نباشد کمتر یا مساوی مقدار زیر است:

$$|H|e^{-\varepsilon m}$$

اثبات. فرض کنید $h_1, h_2, \dots h_k$ تمامی فرضیه‌های درون H باشند که خطای واقعی بیشتر از ε برای c دارند. اگر و فقط اگر حداقل یکی از این k فرضیه با تمامی نمونه‌های آموزشی سازگار باشد فضای ویژه ε-exhausted نخواهد بود. احتمال اینکه فرضیه‌ای که خطای واقعی بیشتر از ε دارد با نمونه‌ای که به صورت اتفاقی انتخاب می‌شود سازگار باشد حداکثر (1-ε) است. بنابراین احتمال اینکه این فرضیه با m نمونه‌ی مستقل سازگار باشد $(1-\varepsilon)^m$ خواهد بود. حال اگر k فرضیه خطایی بیشتر از ε داشته باشند، احتمال اینکه حداقل یکی از این فرضیه‌ها با تمامی m نمونه‌ی آموزشی سازگار باشد حداکثر

$$k(1-\varepsilon)^m$$

است و از آنجایی که $k \leq |H|$، پس این مقدار حداکثر $|H|(1-\varepsilon)^m$ خواهد بود. بالاخره، از رابطه‌ی کلی $0 \leq \varepsilon \leq 1$ داریم که $(1-\varepsilon) \leq e^{-\varepsilon}$. بنابراین،

$$k(1-\varepsilon)^m \leq |H|(1-\varepsilon)^m \leq |H|e^{-\varepsilon m}$$

که قضیه به اثبات می‌رسد.

این قضیه کران بالایی بر حسب تعداد نمونه‌های آموزشی m و حداکثر خطای مجاز ε و اندازه‌ی H برای احتمال اینکه فضای ویژه ε- exhausted نباشد ارائه می‌کند. اما از نظر دیگر، این مرز احتمال اینکه m نمونه‌ی آموزشی در حذف تمامی فرضیه‌های "بد" (فرضیه‌هایی که خطای واقعی بیشتر از ε دارند) در یادگیر سازگار با فضای فرضیه‌ای H موفق نشوند را نشان می‌دهد.

بیایید از این نتیجه برای تعیین تعداد نمونه‌های آموزشی لازم برای کاهش احتمال شکست به زیر حد دلخواه δ استفاده کنیم.

$$|H|e^{-\varepsilon m} \leq \delta \qquad (7.1)$$

با بازنویسی رابطه برای m داریم که

$$m \geq \frac{1}{\varepsilon}\left(\ln|H| + \ln\left(\frac{1}{\delta}\right)\right) \qquad (7.2)$$

به طور خلاصه نامساوی رابطه‌ی ۷,۲ مرزی کلی برای تعداد نمونه‌های آموزشی لازم برای اینکه تمامی یادگیرهای سازگار با موفقیت هر مفهوم هدف درون H را برای مقادیر دلخواه δ و ε یاد بگیرند را مشخص می‌کند. این تعداد نمونه‌ی آموزشی برای تضمین اینکه هر فرضیه‌ی سازگار تقریباً (با احتمال (1-δ) درست) درست (حداکثر با خطای ε) باشد را تعیین می‌کند. توجه دارید که m به صورت خطی با 1/ε و لگاریتمی 1/δ متناسب است. همچنین با نسبت لگاریتمی با اندازه‌ی فضای فرضیه‌ای H نیز متناسب است.

توجه دارید که مرز بالا می‌تواند ذاتاً اغراق‌آمیز باشد. برای مثال، با وجود اینکه احتمال اینکه فضای ویژه باید در بازه‌ی [0,1] قرار بگیرد، اما این مرز با افزایش |H| به صورت خطی افزایش پیدا می‌کند. برای فضای های فرضیه‌ای به اندازه‌ی کافی بزرگ، این مرز می‌تواند به راحتی بزرگ‌تر از یک شود. نتیجه اینکه مرز نامساوی ۷,۲ می‌تواند ذاتاً برای تعداد نمونه‌های آموزشی اغراق‌آمیز باشد. ضعف این مرز معمولاً در جمله‌ی |H| است که از اثبات هنگام جمع احتمالات یک فرضیه‌ی غیرقابل‌قبول در میان تمامی فرضیه‌ها ایجاد می‌شود. در واقع، در بسیاری مواقع مرزی کوچک‌تر فضاهای فرضیه‌ای بی‌نهایت بزرگ را محدود می‌کند. این مرز موضوع قسمت ۷,۴ خواهد بود.

۷,۳,۱ یادگیری agnostic و فرضیه‌های غیر سازگار

رابطه‌ی ۷,۲ از این جهت اهمیت دارد که تعداد نمونه‌های آموزشی لازم برای تضمین اینکه (با احتمال (1-δ)) هر فرضیه‌ی H که خطای آموزشی صفر دارد خطای واقعی حداکثر ε داشته باشد را تعیین می‌کند. متأسفانه اگر H شامل تابع هدف c نباشد، همیشه نمی‌توان فرضیه‌ای پیدا کرد که خطای آموزشی صفر داشته باشد. در چنین حالتی، از یادگیر می‌خواهیم که فرضیه‌ای را خروجی دهد که کمترین خطای ممکن را بر روی نمونه‌های آموزشی داشته باشد. یادگیری که هیچ پیش‌فرضی در مورد تابع هدف نمی‌کند و فقط فرضیه‌ای از H را که کمترین خطای آموزشی دارد را خروجی می‌دهد، یادگیری agnostic نامیده می‌شود، زیرا که هیچ فرض قبلی‌ای برای اینکه آیا C⊆H درست است یا خیر نمی‌کند.

با وجود اینکه رابطه‌ی ۷,۲ بر این فرض که یادگیر فرضیه‌ای با خطای صفر را خروجی می‌دهد پایه‌گذاری شده است، اما مرزی مشابه را می‌توان برای حالت کلی‌تری که یادگیر فرضیه‌ای با خطای آموزشی غیر صفر را خروجی می‌دهد می‌توان به دست آورد. به عبارت دقیق‌تر فرض کنید که D مجموعه‌ی خاصی از نمونه‌ای آموزشی موجود است (البته با $\mathcal{D}$ که توزیع نمونه‌ای است متفاوت است) و $error_D(h)$ خطای نمونه‌ای فرضیه‌ی h بر روی D است. در کل، $error_D(h)$ نسبت اشتباه‌های h بر روی نمونه‌های آموزشی D تعریف شده است. توجه دارید که در حالت کلی $error_D(h)$ با $error_{\mathcal{D}}(h)$ که خطای واقعی بر روی توزیع نمونه‌ای است متفاوت است. حال فرض کنید h_{best} نماد

فرضیه‌ای از H باشد که کمترین خطای نمونه‌ای را بر روی نمونه‌های آموزشی دارد. چه تعداد نمونه‌ی آموزش لازم است تا تضمین شود کـه (بـا احتمال قوی) خطای واقعی $error_{\mathcal{D}}(h_{best})$ کمتر یا مساوی $error_D(h_{best})$ باشد؟ توجه دارید که رابطه‌ی این سؤال با سؤال مطرح شده در قسمت قبلی این است که سؤال قسمت قبلی حالت خاصی از این سؤال بود (حالتی که $error_D(h_{best}) = 0$).

این سؤال را می‌توان با استفاده از تشابه با اثبات قضیه‌ی ۷٫۱ جواب داد (به رابطه‌ی ۷٫۳ مراجعه کنید). یادآوری مرزهـای Hoeffding [۷] (کـه گاهی مرزهای اضافی Hoeffding [۸] نامیده می‌شود.) در اینجا مفید است. مرزهای Hoeffding مشتق بـین احتمـال واقعـی یـک اتفـاق و میزان مشاهده‌ی آن اتفاق در m آزمایش مستقل را بررسی می‌کند. به عبارت دقیق‌تر، این مرزها به m آزمایش مستقل برنولی[۹] اعمال می‌شوند (برای مثال در m پرتاب سکه‌ای با احتمال شیر آمدنی خاص). این کاملاً مشابه تعریف مسئله‌ی ما در تخمین خطای فرضیه در فصـل ۵ اسـت: احتمال اینکه شیر بیاید مشابه احتمال این است که فرضیه یک نمونه‌ی تصادفی را اشتباه دسته‌بندی کند. m پرتاب مستقل سکه مشابه انتخاب m نمونه‌ی مستقل از توزیع نمونه‌ای است. نسبت تعداد شیرها به کل m پرتاب مشابه نسبت دسته‌بندی‌های اشتباه به کل m نمونه‌ی تصادفی است.

مرزهای Hoeffding می‌گوید که اگر خطای نمونه‌ای $error_D(h)$ بر روی مجموعه‌ی D شامل m نمونه‌ی تصادفی باشد، خواهیم داشت که:

$$\Pr[error_{\mathcal{D}}(h) > error_D(h) + \varepsilon] \le e^{-2m\varepsilon^2}$$

این رابطه مرزی برای احتمال اینکه یک فرضیه‌ی دلخواه خطای نمونه‌ای بسیار گمراه‌کننده داشته باشد را به ما می‌دهـد. بـرای اینکـه مطمـئن باشیم که بهترین فرضیه پیدا شده توسط L حداکثر خطایی با این مرز دارد، باید احتمال اینکه هـر فرضـیه از |H| فرضـیه‌های موجـود خطـای بزرگی داشته باشند را در نظر گرفت.

$$\Pr[(\exists h \in H)(error_{\mathcal{D}}(h) > error_D(h) + \varepsilon)] \le |H| e^{-2m\varepsilon^2}$$

اگر این احتمال را δ بنامیم و به دنبال تعداد نمونه‌های لازم برای کمتر بودن δ از مقدار خاصی بگردیم به این رابطه می‌رسیم که:

$$m \ge \frac{1}{2\varepsilon^2}\left(\ln|H| + \ln\left(\frac{1}{\delta}\right)\right) \qquad (7.3)$$

این رابطه تعمیم رابطه‌ی ۷٫۲ برای حالتی است که یادگیر هنوز بهترین فرضیه h∈H را انتخاب می‌کند و خطای نمونه‌ای بهتـرین فرضـیه نیـز می‌تواند غیر صفر باشد. توجه دارید که m با H و 1/δ رابطه‌ای لگاریتمی دارد و همان‌طور که در مشاهده می‌شود که بـه حالـت خـاص‌تر ۷٫۲ می‌رسیم. با این وجود در این حالت کلی‌تر m به جای رابطه‌ی خطی متناسب با مجذور 1/ε است.

[۷] Hoeffding bounds
[۸] Hoeffding additive bounds
[۹] Bernoulli trial

۷،۳،۲ عطف عبارات منطقی PAC-Learnable است

حال که مرزی برای تعیین تعداد نمونه‌های آموزشی کافی برای اینکه بتوان با احتمال خوبی تابع هدف را یاد گرفت به دست آورده‌ایم، از این مرز می‌توانیم برای تعیین پیچیدگی نمونه‌ای و Pac-learnable بودن دسته‌ی خاصی از مفاهیم هدف استفاده کرد.

مجموعه‌ی مفاهیم هدف C را که با عطف عبارات منطقی بیان می‌شود را در نظر بگیرید. یک عبارت منطقی می‌تواند هر متغیر منطقی (مثل Old) یا نقیضش (مثل Old¬) باشد. بنابراین عطف عبارات منطقی مثل توابع هدفی چون "Old ∧ ¬Tall" را نیز شامل می‌شود. آیا C قابل یادگیری PAC است؟ می‌توان نشان داد که پاسخ چنین سؤالی آری است. کافی است ابتدا نشان دهیم هر یادگیر سازگار فقط تعداد چندجمله‌ای‌ای نمونه‌ی آموزشی برای یادگیری هر c در C لازم دارد و الگوریتمی ارائه کنیم که زمانی چندجمله‌ای برای هر نمونه لازم داشته باشد تا مفهوم هدف را یاد بگیرد.

یادگیر L را یک یادگیر سازگاری است در نظر بگیرید که از فضای فرضیه‌ای H که مشابه C است استفاده می‌کند. از رابطه‌ی ۷،۲ می‌توان برای محاسبه‌ی تعداد m نمونه‌ی آموزشی تصادفی کافی تا یادگیر L با احتمال (δ-1) فرضیه‌ای خروجی با ماکزیمم خطای ε بدهد استفاده کرد. برای این کار، لازم است که |H| را که اندازه‌ی فضای فرضیه‌ای مربوطه است را تعیین کرد.

حال فضای فرضیه‌ای H را که بر روی عطف n عبارات منطقی تعریف می‌شود را در نظر بگیرید. اندازه‌ی |H| در این فضای فرضیه‌ای 3^n است. توجه داشته باشید که هر عبارت ممکن است در فرضیه سه حالات داشته باشد: فرضیه آن را شامل می‌شود، فرضیه نقیض آن را شامل می‌شود، فرضیه در مورد آن نظری نداده است. پس اگر n متغیر داشته باشیم می‌توانیم 3^n فرضیه روی آن‌ها تعریف کنیم.

با اضافه کردن $|H| = 3^n$ در رابطه‌ی ۷،۲ مرز پیچیدگی نمونه‌ای یادگیری عطف n عبارت منطقی به صورت زیر به دست می‌آید.

$$m \geq \frac{1}{\varepsilon}\left(nln3 + \ln\left(\frac{1}{\delta}\right)\right) \qquad (7.4)$$

برای مثال اگر یک یادگیر سازگار، یادگیری بخواهد تابع هدفی توصیفی با ۱۰ عبارت و با احتمال درستی بیش از ۹۵ درصد فرضیه‌ای با خطای کمتر از .1 را یاد بگیرد، برای m که تعداد نمونه‌های آموزشی تصادفی لازم برای این کار خواهد بود خواهیم داشت که،

$$.m = \frac{1}{.1}\left(10ln3 + \ln\left(\frac{1}{.05}\right)\right) = 140$$

توجه دارید که m رابطه‌ی خطی و مستقیم با n (تعداد عبارات فضای فرضیه‌ای)، 1/ε و رابطه‌ای لگاریتمی با 1/δ دارد. اما رابطه‌ی m با میزان محاسبات کلی چقدر است؟ البته محاسبات به نوع الگوریتم یادگیری وابسته است. با این وجود، تا زمانی که الگوریتم ما محاسباتی کمتر از چندجمله‌ای برای هر نمونه داشته باشد و کل محاسبات نیز کمتر از چندجمله‌ای کل نمونه‌های آموزشی باشد، مسلماً محاسبات کل نیز کمتر از چندجمله‌ای تعداد نمونه‌ها خواهد بود.

در یادگیری عطف عبارات منطقی، یکی از الگوریتم‌هایی که شرایط لازم را دارد در فصل ۲ مورد بحث قرار گرفت. این الگوریتم Find-S است، که خاص‌ترین فرضیه‌ی سازگار با نمونه‌های آموزشی را محاسبه می‌کند. برای هر نمونه‌ی مثبت آموزشی جدید، الگوریتم اشتراک بین عباراتی فرضیه‌ی فعلی و نمونه‌ی آموزشی جدید را در زمانی با رابطه‌ی خطی با n محاسبه می‌کند. بنابراین، الگوریتم Find-S کلاس مفاهیم عطفی n عبارت منطقی و نقیضشان را به فرم PAC یاد می‌گیرد.

قضیه‌ی ۷,۲ قابلیت یادگیری PAC عطف عبارات منطقی. کلاس C که مجموعه‌ی عطف عبارات منطقی است توسط الگوریتم Find-S با استفاده از H=CFind-S قابل یادگیری PAC است.

اثبات. رابطه‌ی ۷,۴ نشان می‌دهد پیچیدگی نمونه‌ای این کلاس مفاهیم نسبت به $1/\delta$ و $1/\varepsilon$ چندجمله‌ای و از size(c) مستقل است. برای پردازش مرحله به مرحله‌ی هر نمونه‌ی آموزشی، الگوریتم Find-S نیاز به تلاشی متناسب خطی با n و مستقل از $1/\delta$ و $1/\varepsilon$ و size(c) خواهد داشت. بنابراین این کلاس مفاهیم توسط الگوریتم Find-S، قابل یادگیری PAC است.

۷,۳,۳ قابلیت یادگیری PAC دیگر کلاس‌های مفهوم

همان‌طور که در بالا دیدیم، رابطه‌ی ۷,۲ پایه‌ای کلی برای محدود کردن پیچیدگی یادگیری توابع مفهوم کلاس معلوم C ارائه می‌کند. در بالا این رابطه را برای کلاس عطف عبارات منطقی به کار بردیم. به طور مشابه می‌توان نشان داد که بسیاری از کلاس‌های مفهوم پیچیدگی نمونه‌ای چندجمله‌ای دارند. (تمرین ۷,۲)

۷,۳,۳,۱ یادگیرهای بدون بایاس

همه‌ی کلاس‌های مفهوم مرز پیچیدگی نمونه‌ای با رابطه‌ی ۷,۲ محدودی ندارند. برای مثال کلاس مفاهیم بایاس نشده‌ی C را که تمامی مفاهیم قابل‌تعلیم بر روی X را در بر می‌گیرد را در نظر بگیرید. مجموعه‌ی C تمامی مفاهیم هدف قابل‌تعریف همان مجموعه‌ی توانی X، مجموعه‌ی تمامی زیرمجموعه‌های X، خواهد بود که $|C| = 2^{|X|}$. فرض کنید که نمونه‌های درون X با n متغیر منطقی تعریف شوند، بنابراین خواهیم داشت که، $|X| = 2^n$ بنابراین خواهیم داشت که $|C| = 2^{2^n}$. البته برای یادگیری چنین کلاس بایاس نشده‌ی یادگیری، خود یادگیر نیز باید از فضای فرضیه‌ای بدون بایاس استفاده کند H=C. با جایگذاری $|H| = 2^{2^n}$ در رابطه‌ی ۷,۲ پیچیدگی نمونه‌ای برای یادگیر مفاهیم بدون بایاس روی X مشخص می‌شود.

$$m \geq \frac{1}{\varepsilon}\left(2^n ln2 + \ln\left(\frac{1}{\delta}\right)\right) \tag{7.5}$$

بنابراین، این کلاس بدون بایاس از مفاهیم هدف بنا بر رابطه‌ی ۷,۲ پیچیدگی نمونه‌ای نمایی (exponential) در مدل PAC دارد. با وجود اینکه رابطه‌ی ۷,۲ پیچیدگی نمونه‌ای با اغراق برای کلاس مفاهیم بدون بایاس را توانی از n می‌داند اما در حقیقت اثبات می‌شود که این مرز اغراق‌آمیز نیست.

۷,۳,۳,۲ مفاهیم k-term DNF و k-CNF

همچنین می‌توان کلاس‌های مفاهیمی را پیدا کرد که پیچیدگی نمونه‌ای چندجمله‌ای دارند اما با این وجود نمی‌توان آن‌ها را در زمانی چندجمله‌ای یاد گرفت. یکی از مثال‌های جالب چنین کلاس‌هایی، کلاس مفاهیم فرم نرمال فصلی k جمله‌ای (k-term DNF) [۱۰] است. عبارات k-term DNF به فرم $T_1 \vee T_2 \vee \ldots \vee T_k$ هستند که در آن T_i عطفی از n ویژگی منطقی و نقیض‌هایشان است. با این فرض که H=C می تون به راحتی نشان داد که |H| حداکثر 3^{nk} خواهد بود (زیرا که k جمله داریم که هرکدام 3^n حالت دارند). توجه دارید که 3^{nk}

[۱۰] k-term disjunctive normal form

تخمین بالایی از H است زیرا که در شرایطی که $T_i = T_j$ و T_i کلی‌تر از T_j است را دو بار می‌شماریم. با این وجود می‌توان از مرز حداکثری |H| برای پیدا کردن مرز حداکثری پیچیدگی نمونه‌ای استفاده کرد، با جایگذاری در رابطه‌ی ۷,۲ خواهیم داشت،

$$m \geq \frac{1}{\varepsilon}\left(\mathrm{nkln}3 + \ln\left(\frac{1}{\delta}\right)\right) \quad (7.6)$$

این رابطه نشان می‌دهد که پیچیدگی نمونه‌ای k-term DNF چندجمله‌ای از 1/ε، 1/δ، n و k است. با این وجود که پیچیدگی نمونه‌ای از درجه چندجمله‌ای است، پیچیدگی محاسباتی از درجه چندجمله‌ای نیست، زیرا می‌توان نشان داد که این مسئله یادگیری معادل دیگر مسائل یادگیری است که در زمان چندجمله‌ای قابل‌حل نیستند (مگر اینکه RP=NP). بنابراین، با وجود اینکه k-term DNF پیچیدگی چندجمله‌ای دارد، اما پیچیدگی محاسباتی آن برای یادگیری که در آن H=C از درجه‌ی چندجمله‌ای نخواهد بود.

حقیقت جالب در مورد k-term DNF این است که با وجود اینکه این کلاس قابل یادگیری PAC نیست، اما با این حال کلاس مفاهیم بزرگ‌تری وجود دارد که قابل یادگیری PAC است! این از این جهت ممکن است که کلاس‌های مفاهیم بزرگ‌تر پیچیدگی محاسباتی چندجمله‌ای از نمونه‌ها دارند و پیچیدگی نمونه‌ای چندجمله‌ای دارد. این کلاس بزرگ‌تر کلاس نمایش‌های k-CNF است: عطف عبارات با تعداد دلخواه به فرم $T_1 \wedge T_2 \wedge \ldots \wedge T_k$ که در آن T_i فصلی از حداکثر k ویژگی منطقی است. نشان دادن این حکم که k-DNF زیرمجموعه‌ی k-CNF است بسیار ساده است زیرا که می‌توان هر عبارت k-DNF را به سادگی با یک عبارت k-CNF بازنویسی کرد (برعکس این قضیه برقرار نیست). با این وجود که k-CNF نسبت به k-DNF شامل‌تر است، هم پیچیدگی نمونه‌ای چندجمله‌ای و هم پیچیدگی محاسباتی چندجمله‌ای دارد. بنابراین، کلاس مفاهیم k-DNF با الگوریتم یادگیری‌ای که از H=k-CNF استفاده می‌کند قابل یادگیری PAC است. برای بحث دقیق‌تر به (Kearns and Vazirani 1994) مراجعه کنید.

۷,۴ پیچیدگی نمونه‌ای برای فضاهای فرضیه‌ای بیکران

در بخش بالا نشان دادیم که پیچیدگی نمونه‌ای برای یادگیری PAC متناسب با لگاریتم اندازه‌ی فضای فرضیه‌ای است. با وجود اینکه رابطه‌ی ۷,۲ رابطه‌ی بسیار مفیدی است اما دو اشکال در بیان پیچیدگی نمونه‌ای بر اساس |H| وجود دارد. ابتدا اینکه ممکن است به مرزهای ضعیفی ختم گردد (مقدار δ می‌تواند برای مقادیر بزرگ |H| به شکل قابل‌توجهی بزرگ‌تر از یک باشد). دوم اینکه در فضای فرضیه‌ای بیکران به طور کلی نمی‌توان از رابطه‌ی ۷,۲ استفاده کرد.

در اینجا معیار دیگری از پیچیدگی H به نام بعد Vapnik-Chervonenkis H (به اختصار بعد VC H یا VC(H)) را معرفی خواهیم کرد. همان‌طور که در ادامه نیز خواهیم دید، می‌توان مرز پیچیدگی را با معیار VC(H) به جای |H| بیان کرد. در بسیاری از موارد، مرز پیچیدگی بر پایه‌ی VC(H) قوی‌تر از مرز رابطه‌ی ۷,۲ خواهد بود. به علاوه این مرز امکان بررسی پیچیدگی نمونه‌ای بسیاری از فضاهای فرضیه‌ای بیکران را نیز فراهم می‌کند.

۷,٤,۱ خرد کردن مجموعه‌ای از نمونه‌ها

بعد VC پیچیدگی فضای فرضیه‌ای H را نه بر اساس |H| و بلکه بر اساس تعداد نمونه‌های متمایز X که می‌توانند به کلی با H مشخص شوند بیان می‌کند.

برای دقیق‌تر کردن این نمادگذاری، بیایید ابتدا نمادگذاری خرد کردن مجموعه‌ای از نمونه‌ها[۱۱] را مشخص کنیم. زیرمجموعه‌ای از نمونه‌ها مانند $S \subseteq X$ را در نظر بگیرید. برای مثال، شکل ۷,۳ زیرمجموعه‌ای از X شامل سه نمونه را نشان می‌دهد. هر فرضیه‌ی h در H S را به دو مجموعه تقسیم می‌کند؛ این دو مجموعه، مجموعه‌های $\{x \in S|h(x) = 0\}$ و $\{x \in S|h(x) = 0\}$ هستند. با معلوم بودن مجموعه‌ی S می‌توان $2^{|S|}$ تقسیم دوتایی مختلفی که اعضای H ممکن بعضی از آن‌ها را نتوانند ایجاد کنند نوشت. زمانی می‌گوییم H، S را خرد می‌کند که برای هر یک از تقسیم‌های دوتایی S را بتوان با فرضیه‌ای از H نمایش داد.

تعریف. مجموعه‌ی نمونه‌های S با فضای فرضیه‌ای H خرد می‌شود اگر و فقط اگر برای هر تقسیم دوتایی S فرضیه‌ای سازگار وجود داشته باشد.

شکل ۷,۳ مجموعه‌ای از سه نمونه را نشان می‌دهد که توسط فضای فرضیه‌ای خرد می‌شود. توجه دارید که برای هر یک از 2^3 تقسیم دوتایی این سه نمونه فرضیه‌ای وجود دارد.

توجه دارید که اگر مجموعه‌ای از نمونه‌ها توسط یک فضای فرضیه‌ای خرد نشود، بدین معناست که فرضیه‌ای (تقسیم دوتایی‌ای) بر روی نمونه‌ها وجود دارد که نمی‌توان آن را با فضای فرضیه‌ای نشان داد. قدرت خرد کردن یک فضای فرضیه‌ای برای مجموعه‌ای از نمونه‌ها معیاری از قدرت نمایش این فضای فرضیه‌ای برای نمایش مفاهیم تعریف شده بر روی این مجموعه از نمونه‌هاست.

۷,٤,۲ بعد Vapnik-Chervonenkis

قدرت خرد کردن مجموعه‌ای از نمونه‌ها رابطه‌ی نزدیکی با بایاس استقرایی یک فضای فرضیه‌ای دارد. با توجه به آنچه در فصل ۲ گفته شد، فضای فرضیه‌ای بدون بایاس فضای فرضیه‌ای است که می‌تواند تمامی مفاهیم (تقسیم‌های دوتایی) قابل‌تعریف روی فضای نمونه‌ی را نمایش بدهد. اما اگر H نتواند X را خرد کند، اما در مقابل بتواند زیرمجموعه‌ی بزرگی از X را خرد کند چه؟ به نظر می‌رسد اینکه هر قدر زیرمجموعه‌ی خردشده‌ی X بزرگ‌تر باشد، H نیز شامل‌تر خواهد بود. بعد VC ی H به طور دقیق‌تر معیار زیر است.

تعریف. بعد Vapnik-Chervonenkis، VC(H)، برای فضای فرضیه‌ای H که بر روی فضای نمونه‌ای X تعریف شده اندازه‌ی بزرگ‌ترین زیرمجموعه‌ی کران‌دار X است که با H خرد می‌شود. اگر H بتواند هر زیرمجموعه‌ی دلخواه X را خرد کند خواهیم داشت، VC(H)≡∞.

توجه دارید که برای تمامی H های کران‌دار داریم $VC(H) \leq \log_2 |H|$. برای درک این رابطه فرض کنید داریم $VC(H) = d$. پس H به 2^d فرضیه‌ی محض برای خرد کردن d نمونه احتیاج خواهد داشت، پس $2^d \leq |H|$ و $d = VC(H) \leq \log_2 |H|$.

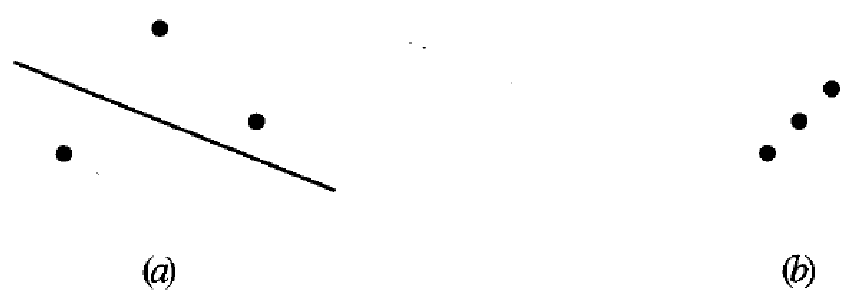
---

[۱۱] shattering a set of instances

۷,٤,۲,۱ چندین مثال

برای پیدا کردن درکی از VC(H)، چند نمونه فضای فرضیه‌ای را در نظر بگیرید. برای شروع، فرض کنید که X مجموعه‌ی اعداد حقیقی $X=\mathfrak{R}$است (که قد افراد را توصیف می‌کند) و H مجموعه‌ی بازه‌های اعداد حقیقی است. به عبارت دیگر H مجموعه‌ی فرضیه‌هایی است که به فرم $a<x<b$ بیان می‌شوند و a و b نیز اعداد ثابت حقیقی‌اند. VC(H) در اینجا چیست؟ برای جواب به این سؤال، باید بزرگ‌ترین زیرمجموعه‌ی X را پیدا کنیم که با H خرد شود. زیرمجموعه‌ی دو عضوی خاصی از اعداد حقیقی مثل S={3.1,5.7} را در نظر بگیرید. آیا می‌توان S را با H خرد کرد؟ بله، برای مثال چهار فرضیه‌ی $(2<x<1)$، $(4<x<1)$، $(4<x<7)$ و $(1<x<7)$ این کار را انجام می‌دهند. این چهار مجموعه با هم هر یک از تقسیم‌های دو عضوی S را، مثل دربر داشتن هیچ‌کدام، یکی و هر دو نمونه، را نمایش می‌دهند. از آنجایی که مجموعه‌ای دو عضوی پیدا کردیم که با H خرد شود پس VC(H) حداقل دو خواهد بود. اما آیا مجموعه‌ای با اندازه‌ی سه وجود دارد که H آن را خرد کند؟ مجموعه‌ی $S=\{x_0,x_1,x_2\}$ با سه نمونه‌ی دلخواه را در نظر بگیرید. بدون از دست دادن کلیت فرض کنیم که $x_0 < x_1 < x_2$. واضح است که این زیرمجموعه را نمی‌توان خرد کرد، زیرا که تقسیم دو عضوی‌ای که x_0 و x_2 را در برگرفته و x_1 را در بر نگیرد را نمی‌توان با یک فرضیه نشان داد. بنابراین، هیچ زیرمجموعه‌ی سه عضوی را نمی‌توان خرد کرد و VC(H)=2. توجه دارید که در اینجا H بیکران و VC(H) کران‌دار است.

فرض کنید که X مجموعه نقاط روی صفحه‌ی x-y باشد (شکل ۷,۴). فرض کنیم H تمام سطوح تصمیم‌گیری خطی بر روی این صفحه است. به عبارت دیگر، H فضای فرضیه‌ای متناسب با یک واحد پرسپترون با دو ورودی است (برای بحث کلی‌تر به فصل ۴ مراجعه کنید). بعد VC ی H چیست؟ درک اینکه هر دو نقطه در فضا را می‌توان با فضای فرضیه‌ای H خرد کرد بسیار آسان است، زیرا که چهار خط پیدا می‌شود که هیچ‌کدام، یکی و یا هردو نمونه را در بر بگیرند؟ اما در مورد مجموعه‌های سه نقطه‌ای چه؟ تا زمانی که نقاط بر روی یک خط نیستند، 2^3 خط پیدا خواهیم کرد که مجموعه را خرد کند. البته مجموعه‌ی سه نقطه‌ی روی یک خط را نمی‌توان خرد کرد (به همان دلیل که نمی‌توانستیم در مثال قبل مجموعه‌ی سه نقطه‌ای از روی خط حقیقی را خرد کنیم). در چنین شرایطی VC(H) چند است، دو یا سه؟ حداقل سه است. تعریف بعد VC می‌گوید که اگر بتوانیم مجموعه‌ای d عضوی را پیدا کنیم که بتوان آن را با H خرد کرد داریم که $VC(H) \geq d$. برای نشان دادن اینکه $VC(H) < d$ باید نشان دهیم که مجموعه‌ی d عضوی را نمی‌توان با H خرد کرد. در این مثال، هیچ مجموعه‌ی چهار عضوی را نمی‌توان خرد کرد پس VC(H)=3. به طور کلی می‌توان نشان داد که بعد VC ی سطوح تصمیم خطی در فضای r بعدی (بعد VC پرسپترونی با r ورودی) r+1 است.



شکل ۷,۴ بعد VC ی سطوح تصمیم خطی در صفحه‌ی x-y ۳ است.
(a) مجموعه‌ای از سه نقطه که با سطوح تصمیم خطی خرد می‌شود. (b) مجموعه‌ای از سه نقطه که نمی‌توان آن را با سطوح تصمیم خطی خرد کرد.

به عنوان مثال آخر، فرض کنید که هر نمونه از X با عطفی از سه عبارت منطقی، و هر فرضیه‌ی H عطف حداکثر سه عبارت منطقی نمونه‌ها باشد. VC(H) چیست؟ می‌توان نشان داد که این مقدار حداقل ۳ است. هر یک از نمونه‌ها را با رشته‌ای سه بیتی متناسب با عبارات l_1، l_2 و l_3 نشان می‌دهیم. مجموعه‌ی سه نمونه‌ای زیر را در نظر بگیرید:

$instance_1: 100$

$instance_2: 010$

$instance_3: 001$

این مجموعه‌ی سه عضوی از نمونه‌ها را می‌توان با H خرد کرد، زیرا که برای هر تقسیم دوتایی می‌توان به فرم زیر فرضیه‌ای ساخت: اگر فرضیه‌ای نمونه‌ی $instance_i$ را در بر نمی‌گیرد $\neg l_i$ را به فرضیه اضافه کن. برای مثال فرضیه‌ای را در نظر بگیرید که $instance_2$ را دربرگرفته اما $instance_1$ و $instance_3$ را دربر نمی‌گیرد. از فرضیه‌ی $\neg l_1 \wedge \neg l_2$ برای این حالت استفاده خواهیم کرد. این بحث را می‌توان به سادگی از ۳ ویژگی به n ویژگی تعمیم داد. در واقع، VC(H) در این حالت دقیقاً n است، اما نشان دادن این کار کمی سخت‌تر است زیرا باید نشان دهیم که مجموعه‌ی n+1 عضوی‌ای وجود ندارد که با H خرد شود.

۷,٤,۳ پیچیدگی نمونه‌ای و بعد VC

در قسمت‌های قبلی سؤال "چه تعداد نمونه‌ی تصادفی برای تخمین یکی از فرضیه‌های C کافی است؟" (چه تعداد نمونه‌ی آموزشی برای اینکه با احتمال (1-δ) فضای ویژه ε-exhaust باشد؟) را بررسی کردیم. با استفاده از VC(H) به عنوان معیاری برای پیچیدگی H چگونه می‌توان جوابی دیگر برای این سؤال پیدا کرد، مشابه آنچه پیش‌تر با مرز رابطه‌ی ۷,۲ بیان کردیم. این مرز جدید به فرم زیر است (به (Blumer et al. 1989 مراجعه کنید)

$$m \geq \frac{1}{\varepsilon}\left(4\log_2\left(\frac{2}{\delta}\right) + 8VC(H)\log_2\left(\frac{13}{\varepsilon}\right)\right) \qquad (7.7)$$

توجه دارید که مشابه رابطه‌ی ۷,۲ تعداد نمونه‌های لازم m با لگاریتم 1/δ متناسب است. اما به جای رابطه‌ی خطی با لگاریتم برابر رابطه خطی با 1/ε متناسب است. قابل‌توجه است که، جمله‌ی ln|H| که در مرز قبلی بود با معیار جایگزین پیچیدگی فضای فرضیه‌ای، VC(H)، جایگزین شده است. (توجه دارید که $VC(H) \leq \log_2|H|$).

رابطه‌ی ۷,۷ کران بالایی برای تعداد نمونه‌های آموزشی لازم برای اینکه هر فرضیه‌ی C را به طور PAC با ε و δ دلخواه یاد بگیریم را معلوم می‌کند. پیدا کردن کران پایین برای این تعداد با استفاده از قضیه‌ی زیر امکان‌پذیر است (به (Ehrenfeucht et al. 1989) مراجعه کنید).

قضیه‌ی ۷,۳ کران پایین پیچیدگی نمونه‌ای. کلاس مفاهیم دلخواه C که برای آن داریم $VC(C) \geq 2$، یادگیر دلخواه L و $0 < \varepsilon < \frac{1}{8}$ و $0 < \delta < \frac{1}{100}$ را در نظر بگیرید. توزیعی مثل $\mathcal{D}$ و مفهوم هدفی مثل c در C وجود دارد که اگر L کمتر از

$$max\left[\frac{1}{\varepsilon}\log\left(\frac{1}{\delta}\right), \frac{VC(C)-1}{32\varepsilon}\right]$$

تعداد نمونه را مشاهده کرده باشد، با احتمال حداقل δ، L فرضیه‌ای را خروجی می‌دهد که $error_{\mathcal{D}}(h) > \varepsilon$.

این قضیه نشان می‌دهد که اگر تعداد نمونه‌های آموزشی بسیار کم باشد، هیچ یادگیری نمی‌تواند تمامی مفاهیم هدف C غیربـدیهی را بـه طـور PAC یاد بگیرد. بنابراین، این قضیه کران پایینی بر روی تعداد نمونه‌های آموزشی برای یادگیری موفق را ارائه می‌کند، این مرز کامـل‌کننـده‌ی کران بالای ذکر شده برای تعداد کافی نمونه‌های آموزشی است. توجه دارید که این کران پایین با پیچیدگی کلاس مفـاهیم C بیـان می‌شـود، در حالی که کران بالا با H تعیین می‌شد. (چرا؟)

این کران پایین نشان می‌دهد که کران بالای نامساوی ۷٫۷ به اندازه‌ی کافی محکم است. هر دو کران با $1/\delta$ رابطه‌ی لگاریتمی و با $VC(H)$ رابطه‌ی خطی دارند. تنها تفاوت‌های باقی‌مانده در این دو کران وابستگی کران بالا به $\log(1/\varepsilon)$ است.

۷٫٤٫٤ بعد VC برای شبکه‌های عصبی

با در نظر داشتن بحث شبکه‌های عصبی مصنوعی از فصل ۴، تعیین بعد VC شبکه‌ای از واحدهای مرتبط، مثل شبکه‌های عصبی تک سویه که توسط فرایند backpropagation آموزش داده می‌شوند، جالب خواهد بود. این بخش نتیجه‌ی کلی‌ای از محاسبه‌ی بعـد VC شـبکه‌های بدون دور را بر اساس ساختار شبکه و بعد VC خود واحدها را ارائه می‌کند. این بعد VC را می‌توان برای محدود کردن نمونه‌های آموزشی لـازم برای یادگیری تقریباً درست شبکه‌ی تک سویه برای مقادیر دلخواه ε و δ به کاربرد. می‌توانید در اولین مطالعه‌ی کتاب ایـن بخـش را بـدون از دست دادن پیوستگی مطلب نخوانید.

شبکه‌ی G متشکل از واحدها با گراف بدون دور را در نظر بگیرید. یک گراف جهت‌دار[۱۲] بدون دور[۱۳] گرافی اسـت کـه یال‌هـایش جهـت دارنـد (واحدها ورودی و خروجی هستند) و دور ندارد. گراف لایه‌ای[۱۴] گرافی است که گره‌هایش را بتوان به صورتی تقسیم‌بندی کرد که تمامی یال‌های جهت‌دار خروجی از گره‌های لایه l به گره‌های لایه‌ی l+1 بروند. گراف شبکه‌ی عصبی تک سویه در فصل ۴، مثالی از چنین گراف‌های جهت‌دار لایه‌ای بدون دور است.

ثابت می‌شود که می‌توان بعد VC چنین شبکه‌هایی را بر اساس ساختارشان و بعد VC واحدهای اولیه‌ی سازنده‌شان محدود کرد. بـرای فرمولـه کردن این حقیقت، باید ابتدا چندین عبارت دیگر را تعریف کنیم. بیایید فرض کنیم که n تعداد ورودی‌های شبکه‌ی G است و ایـن شـبکه تنهـا یک خروجی دارد. فرض کنیم که واحدهای داخلی G، N_i (هر واحدی که ورودی نباشـد) حـداکثر r ورودی داشـته و از تـابعی منطقـی مقـدار $c_i: \mathfrak{R}^r \rightarrow \{0,1\}$ از کلاس توابع C استفاده کند. برای مثال اگر واحدهای داخلی پرسپترون باشـند C کلـاس توابـع خطـی مقـدار آسـتانه‌ای تعریف شده بر روی $\mathfrak{R}^r$ خواهد بود.

حال می‌توانیم ترکیب G[۱۵] ی C را به عنوان کلاس تمام توابعی که شبکه G می‌تواند به کار ببرد با این فرض که هر واحد شبکه‌ی G یکـی از توابع کلاس C را مورد استفاده قرار دهد تعریف کرد. به طور خلاصه، ترکیب G ی C فضای فرضیه‌ای است که توسط شـبکه‌ی G قابل‌نمـایش است.

قضیه‌ی زیر بعد VC ترکیب G ی C را بر اساس بعد VC ی C و ساختار G محدود می‌کند.

---

[۱۲] directed
[۱۳] acyclic
[۱۴] layered graph
[۱۵] G-composition

قضیه‌ی ۷،۴. بعد VC شبکه‌های جهت‌دار لایه‌ای بدون دور. (برای اطلاعات بیشتر به (Kearns and Vazirani 1994) مراجعه کنید). فرض کنید G گراف جهت‌دار لایه‌ای بدون دوری با n گره ورودی و $s \geq 2$ گره داخلی با حداکثر r ورودی است و C نیز کلاس مفاهیم روی $\mathfrak{R}^r$ با بعد VC، d است که متناسب با دسته توابع قابل توصیف توسط هر یک از s گره داخلی است. اگر C_G ترکیب G ی C متناسب با دسته توابع قابل توصیف با G باشد، داریم که $VC(C_G) \leq 2ds\log(es)$ که در این رابطه e پایه‌ی لگاریتم طبیعی (عدد نپر) است.

توجه دارید که این مرز بعد VC شبکه‌ی G رابطه‌ی خطی با بعد VC ی d تک واحدهایش و رابطه‌ی لگاریتمی با s، تعداد واحدهای آستانه‌ی شبکه دارد.

شبکه‌های لایه‌ای بدون دوری را در نظر بگیرید که از گره‌های پرسپترون تشکیل یافته‌اند. با توجه به آنچه در فصل ۴ گفته شد، پرسپترون r ورودی از سطوح تصمیم خطی برای نمایش توابع منطقی بر روی $\mathfrak{R}^n$ استفاده می‌کند. همان‌طور که در بخش ۷،۴،۲،۱ نیز گفته شد، بعد VC ی سطوح تصمیم خطی ی $\mathfrak{R}^n$، r+1 است. بنابراین، یک تک پرسپترون با r ورودی بعد VC ی r+1 خواهد داشت. از این حقیقت می‌توان به همراه قضیه‌ی بالا برای محدود کردن بعد VC شبکه‌ای لایه‌ای شامل s پرسپترون هر کدام با r ورودی استفاده کرد،

$$\mathrm{VC}\left(\mathrm{C}_{\mathrm{G}}^{\mathrm{perceprtons}}\right) \leq 2(\mathrm{r}+1)\mathrm{s}\log(\mathrm{es})$$

حال می‌توان تعداد m نمونه‌ی آموزشی کافی برای یادگیری (با احتمال حداقل (δ-1)) هر مفهوم هدف $\mathrm{C}_{\mathrm{G}}^{\mathrm{perceprtons}}$ را با خطای ε مشخص کرد. با جایگذاری رابطه‌ی بالا برای VC شبکه در رابطه‌ی ۷،۷ خواهیم داشت،

$$\begin{aligned} m &\geq \frac{1}{\varepsilon}\left(4\log\left(\frac{2}{\delta}\right) + 8\mathrm{VC(H)}\log\left(\frac{13}{\delta}\right)\right) \\ &\geq \frac{1}{\varepsilon}\left(4\log\left(\frac{2}{\delta}\right) + 16(\mathrm{r}+1)\mathrm{s}\log(\mathrm{es})\log\left(\frac{13}{\delta}\right)\right) \end{aligned} \tag{7.8}$$

همان‌طور که با این مثال شبکه‌ی پرسپترن نشاند داده شد، قضیه‌ی بالا از این نظر جالب است که متدی کلی برای محدود کردن بعد VC ی شبکه‌ای بدون دور و لایه‌ای از واحدها را بر اساس ساختار شبکه و بعد VC تک واحد سازنده محدود می‌کنیم. متأسفانه نتایج بالا مستقیماً در مورد شبکه‌هایی که با Backpropagation آموزش داده می‌شوند صادق نیست، به دو دلیل. اول اینکه این نتایج برای شبکه‌های پرسپترون، و نه واحدهای سیگموید که در backpropagation مورد استفاده است، نتیجه‌گیری شده است. با این وجود، توجه دارید که بعد VC ی واحدهای سیگموید حداقل به اندازه‌ی بعد VC ی واحدهای پرسپترون است، زیرا که واحد سیگموید می‌تواند پرسپترون را تا حد دلخواه با افزایش وزن‌ها تخمین بزند. بنابراین، مرز بالا برای m حداقل مرز ممکن برای شبکه‌های بدون دور واحدهای سیگموید است. مشکل دوم تعمیم نتیجه‌ی بالا این است که Backpropagation از شبکه‌ای با وزن‌های غیر صفر کار خود را شروع کرده و با تغییر وزن‌ها به فرضیه‌ی قابل‌قبول می‌رسد. بنابراین، Backpropagation با معیار توقف cross-validation بایاسی استقرایی با ترجیح شبکه‌هایی با وزن‌های کوچک‌تر دارد. این بایاس استقرایی که به طور مؤثری VC را کاهش می‌دهد در بررسی بالا در نظر گرفته نشده است.

۷,۵ مدل یادگیری مرز خطا

با وجود اینکه ما بیشتر بر روی مدل یادگیری PAC تمرکز کردیم، تئوری یادگیری محاسباتی تعریف مسئله‌های دیگر و دیگر سؤالات را در نظر دربر می‌گیرد. تعریف مسئله‌های یادگیری مختلفی که مورد مطالعه قرار گرفته است در نحوه‌ی ایجاد نمونه‌های یادگیری (مشاهده‌ی سوم شخص[۱۶] نمونه‌های تصادفی، انتخاب آزمایش توسط یادگیر)، نویز داده‌ها (با خطا یا بدون خطا)، تعریف موفق (مفهوم هدف باید دقیقاً یاد گرفته شود یا اینکه تقریباً یا با احتمال خاصی یاد گرفته شود)، فرض‌های یادگیر (شامل توزیع نمونه‌ای و اینکه $C \subseteq H$) و معیاری که با آن یادگیر ارزیابی می‌شود (تعداد نمونه‌های آموزشی، تعداد اشتباه‌ها، زمان کل یادگیری) متفاوت‌اند.

در این بخش به مدل یادگیری مرز خطا، که در آن یادگیر با تعداد اشتباه‌هایش قبل از همگرایی به فرضیه‌ی درست ارزیابی می‌شود خواهیم پرداخت. مشابه تعریف مسئله‌ی PAC، فرض می‌کنیم یادگیر سری‌ای از نمونه‌های آموزشی را دریافت می‌کند. با این وجود، در اینجا می‌خواهیم یادگیر قبل از دریافت هر نمونه‌ی x مقدار تابع هدف c(x) را (قبل از معلوم شدن مقدار درست هدف توسط آموزش‌دهنده) پیش‌بینی کند. سؤال مطرح این است که "یادگیر قبل از یادگیری مفهوم هدف چه تعداد پیش‌بینی اشتباه خواهد کرد؟" اهمیت این سؤال در کاربرد عملی است، زیرا که یادگیری باید زمانی که سیستم در حال استفاده واقعی است انجام شود، نه در مرحله‌ی آموزشی مجزا. برای مثال، اگر سیستم برای یادگیری پیش‌بینی اینکه چه پرداخت‌های credit cart ی باید ثبت شود و چه پرداخت‌هایی تقلبی هستند بر اساس اطلاعاتی که حین استفاده از سیستم جمع‌آوری می‌کند طراحی می‌شود، بنابراین علاقه خواهیم داشته که تعداد اشتباهات قبل از همگرایی به تابع هدف مینیمم شود. در اینجا تعداد کل اشتباهات می‌تواند اهمیت بیشتری نسبت به تعداد کل نمونه‌های آموزشی داشته باشد.

این مسئله‌ی یادگیری مرز خطا را می‌توان در شرایط خاص مختلفی مورد مطالعه قرار داد. برای مثال ممکن است تعداد اشتباهات قبل از یادگیری PAC تابع هدف را بشماریم. اما در مثال‌های زیر ما تعداد اشتباه‌ها قبل از اینکه یادگیر مفهوم هدف را دقیق یاد بگیرید را در نظر می‌گیریم. یادگیری مفهوم هدف به طور دقیق بدین معناست که به فرضیه‌ای میل کنیم که $(\forall x) h(x) = c(x)$.

۷,۵,۱ مرز خطای الگوریتم **Find-S**

برای تصور، دوباره فضای فرضیه‌ای H عطف n عبارت منطقی $l_1 \dots l_n$ و نقیض‌هایشان را در نظر بگیرید (برای مثال Rich∧¬Handsome). الگوریتم Find-S که خاص‌ترین فرضیه سازگار با نمونه‌های آموزشی را محاسبه می‌کرد را از فصل ۲ به یاد بیاورید. یکی از سرراست‌ترین پیاده‌سازی الگوریتم Find-S برای فضای فرضیه‌ای H در زیر آمده:

Find-S:

فرضیه‌ی h را با خاص‌ترین فرضیه $l_1 \wedge \neg l_1 \wedge l_2 \wedge \neg l_2 \dots l_n \wedge \neg l_n$ مقداردهی اولیه کن.

برای هر نمونه‌ی مثبت x

هر عبارت h را که توسط x راضی نمی‌شد را حذف کن.

فرضیه‌ی h را خروجی بده.

---

[۱۶] passive

Find-S به طور حدی به فرضیه‌ای میل می‌کند که خطایی نخواهد داشت، به شرطی که C⊆H باشد و داده‌های آموزشی نیز بدون خطا باشند. Find-S با خاص‌ترین فرضیه (فرضیه‌ای که تمامی نمونه‌ها را منفی دسته‌بندی می‌کند) شروع می‌کند، سپس به صورت پلکانی این فرضیه را در مواقع لازم برای پوشاندن نمونه‌های آموزشی مثبت کلی‌تر می‌کند. برای نمایش فرضیه‌ای استفاده شده در اینجا، این کلی سازی با حذف عبارات راضی نشده خواهد بود.

آیا می‌توان اثبات کرد که تعداد اشتباهات Find-S قبل از یادگیری دقیق مفهوم هدف c کمتر از تعداد خاصی است؟ جواب بلی است. برای درک این، توجه کنید که اگر داشته باشیم c∈H، آنگاه هیچ‌گاه Find-S نمونه‌ی منفی‌ای را مثبت دسته‌بندی نخواهد کرد. دلیل این است که فرضیه‌ی فعلی h همیشه حداقل به اندازه‌ی مفهوم هدف c خاص است. بنابراین، برای محاسبه‌ی تعداد اشتباهات، فقط باید اشتباهاتی را که نمونه‌ی مثبت، منفی دسته‌بندی می‌شود را بشماریم. این گونه اشتباهات قبل از یادگیری کامل c چند بار اتفاق خواهند افتاد؟ اولین نمونه‌ی مثبت ارائه شده به Find-S را در نظر بگیرید. یادگیر در دسته‌بندی این نمونه دقیقاً یک اشتباه انجام خواهد داد، زیرا که فرضیه‌ی اولیه‌ی یادگیر تمامی نمونه‌ها را منفی دسته‌بندی می‌کند. با این وجود، نتیجه این خواهد بود که نصف 2n عبارت فرضیه‌ی اولیه حذف خواهد شد و n عبارت باقی خواهد ماند. برای هر نمونه‌ی مثبت بعدی که به اشتباه منفی دسته‌بندی می‌شود حداقل یکی از این n عبارت از فرضیه‌ی h حذف خواهد شد. بنابراین، تعداد کل اشتباهات حداکثر n+1 خواهد بود. این تعداد اشتباه در بدترین حالت رخ خواهد داد، یعنی از یادگیر بخواهیم کلی‌ترین مفهوم هدف $(\forall x)c(x)=1$ را یاد بگیرد بدترین سری نمونه‌ها، یعنی نمونه‌هایی که در هر بار اشتباه فقط یک عبارت را حذف می‌کنند به یادگیر داده شود.

۷,۵,۲ مرز خطای الگوریتم Halving

به عنوان مثال دوم، الگوریتمی را در نظر بگیرید که با نگه داشتن توصیفی از فضای ویژه یاد می‌گیرد، و پلکانی در این فضای ویژه با برخورد با نمونه‌های جدید تجدیدنظر می‌کند. الگوریتم‌های Candidate-Elimination و List-Then-Eliminate در فصل ۲ چنین الگوریتم‌هایی هستند. در این بخش مرز بدترین حالت ممکن[۱۷] را روی تعداد اشتباه‌هایی که چنین یادگیری انجام می‌دهد را برای فضای فرضیه‌ای محدود H، دوباره با فرض اینکه تابع هدف را باید دقیقاً یاد بگیریم، محاسبه می‌کنیم.

برای بررسی تعداد اشتباهات حین یادگیری، ابتدا بادید تعیین کنیم که یادگیر دقیقاً چگونه دسته‌بندی نمونه‌ی جدید را پیش‌بینی می‌کند. بیایید فرض کنیم که این پیش‌بینی با رأی‌گیری در بین فرضیه‌های فضای ویژه فعلی انجام می‌گیرد. اگر اکثر فرضیه‌های فضای ویژه نمونه‌ی جدید را مثبت دسته‌بندی کنند، بنابراین این پیش‌بینی، پیش‌بینی یادگیر نیز خواهد بود. در غیر این صورت پیش‌بینی یادگیر منفی خواهد بود.

این ترکیب یادگیری فضای ویژه، به همراه رأی اکثریت برای پیش‌بینی‌های بعدی، گاهی الگوریتم Halving نامیده می‌شود. ماکزیمم تعداد اشتباهات الگوریتم Halving برای H محدود دلخواه قبل از یادگیری مفهوم هدف چیست؟ توجه دارید که یادگیری "دقیق"[۱۸] مفهوم هدف، متناسب با رسیدن به حالتی است که فضای ویژه فقط شامل یک فرضیه شود. (مشابه معمول، فرض می‌کنیم که مفهوم هدف c در H وجود دارد).

---

[۱۷] wose-case bound

[۱۸] exact

برای به دست آوردن مرز خطا، توجه داشته باشید که الگوریتم Halving فقط زمانی اشتباه می‌کند که اکثریت فضای ویژه‌اش نمونه‌ی جدید را اشتباه دسته‌بندی کنند. در چنین شرایطی، هنگامی که دسته‌بندی جدید برای یادگیر آشکار می‌شود، اندازه‌ی فضای ویژه حداقل به نصف اندازه‌ی فعلی‌اش کاهش می‌یابد (فقط فرضیه‌هایی که اقلیت بودند باقی می‌مانند). با معلوم بودن اینکه با هر اشتباه اندازه‌ی فضای ویژه حداقل به نصف کاهش می‌یابد و با دانستن اینکه فضای ویژه‌ی اولیه فقط |H| عضو داشته، حداکثر اشتباهات ممکن قبل از اینکه فضای ویژه فقط یک عضو داشته باشد $\log_2|H|$ است. در واقع می‌توان نشان داد که مقدار این مرز $\lfloor\log_2|H|\rfloor$ است. برای مثال، فرض کنید که |H|=7 است. اولین اشتباه اندازه‌ی |H| را به ۳ و دومین اشتباه اندازه‌ی آن را به ۱ کاهش خواهد داد.

توجه دارید که مرز $\lfloor\log_2|H|\rfloor$ مرز بدترین حالت است، و ممکن است الگوریتم Halving بدون هیچ اشتباهی مفهوم هدف را یاد بگیرد. دلیل این حقیقت این است که حتی زمانی که رأی اکثریت درست است، الگوریتم فرضیه‌های اشتباه، اقلیت، را حذف خواهد کرد. اگر چنین اتفاقی پیاپی در طول سری آموزش رخ بدهد، بنابراین ممکن است بدون هیچ اشتباهی فضای ویژه به یک عضو کاهش داده شود.

یکی از تغییرات جالب الگوریتم Halving قائل شدن وزن برای رأی فرضیه‌هاست. فصل ۶ دسته‌بندی کننده‌ی بهینه‌ی بیز، که از رأی‌گیری وزن‌دار میان فرضیه‌ها استفاده می‌کند را معرفی می‌کند. در دسته‌بندی کننده‌ی بهینه‌ی بیز، وزن نسبت داده شده به هر فرضیه احتمال ثانویه‌ی تخمینی توصیف مفهوم هدف به شرط مشاهده‌ی داده‌های آموزشی است. در ادامه‌ی این بخش الگوریتمی متفاوت بر پایه‌ی رأی‌گیری وزن‌دار دیگری را به نام Weighted-Majority معرفی خواهیم کرد.

۷,۵,۳ مرزهای خطای بهینه

بررسی بالا مرز بدترین حالت اشتباه را برای دو الگوریتم خاص، Find-S و Candidate-Elimination، تعیین کرد. تعیین اینکه مرز بهینه‌ی خطا برای کلاس مفاهیم دلخواه C با فرض اینکه H=C جالب خواهد بود. منظور از مرز خطای بهینه، کمترین مرز خطای بدترین حالت برای تمامی الگوریتم‌های یادگیری ممکن است. به عبارت دقیق‌تر، برای یادگیری الگوریتم A و مفهوم هدف c $M_A(c)$ را ماکزیمم خطای A در تمامی سری‌های ممکن نمونه‌های آموزشی برای یادگیری دقیق c تعریف می‌کنیم. حال برای هر کلاس فرضیه‌ای غیر تهی تعریف می‌کنیم که $M_A(C) \equiv \max_{c\in C} M_A(c)$. توجه دارید که در بالا نشان دادیم که $M_{Find-S}(C) = n+1$ با این فرض که C کلاس مفاهیم عطف حداکثر n ویژگی منطقی باشد. همچنین نشان دادیم که $M_{Halving}(C) \le \log_2(|C|)$ برای تمامی کلاس‌های مفهوم C است.

مرز خطای بهینه را برای کلاس مفاهیم C به فرم زیر تعریف می‌کنیم.

تعریف. اگر C کلاس غیرتهی دلخواهی باشد، مرز بهینه‌ی خطای C، که با Opt(C) نمایش داده می‌شود، مینیمم روی تمام الگوریتم‌های ممکن A $M_A(C)$ است.

$$Opt(C) \equiv \min_{A\in learnig\ algorithms} M_A(C)$$

به طور غیررسمی، این تعریف Opt(C) را تعداد اشتباهات سخت‌ترین C با استفاده از سخت‌ترین سری نمونه‌های آموزشی برای بهترین الگوریتم یادگیری تعریف می‌کند. (Littlestone 1987) نشان می‌دهد که برای هر کلاس مفهوم C، رابطه‌ای جالب میان مرز خطای بهینه‌ی C و مرز خطای الگوریتم Halving و بعد VC ی C وجود دارد،

$$\mathrm{VC} \le \mathrm{Opt(C)} \le \mathrm{M_{Halving}(C)} \le \log_2(|C|)$$

علاوه بر این کلاس‌های مفاهیمی وجود دارد که این چهار کمیت برای آن‌ها دقیقاً مساوی است. یکی از چنین کلاس‌های مفاهیم کلاس مجموعه‌ی توانی $\mathrm{C_p}$ برای مجموعه‌ی محدود X است. در چنین شرایطی $VC(\mathrm{C_p}) = |X| = \log_2(|\mathrm{C_p}|)$. بنابراین تمامی چهار کمیت بالا برابرند. (Littlestone 1987) نمونه‌هایی از دیگر کلاس‌های مفاهیمی که VC(C) مطلقاً کمتر از Opt(C) و Opt(C) مطلقاً از $\mathrm{M_{Halving}}(\mathrm{C})$ کمتر است ارائه می‌کند.

۷,۵,۴ الگوریتم Weighted-Majority

در این قسمت تعمیمی از الگوریتم Halving را به نام الگوریتم Weighted-Majority بررسی می‌کنیم. الگوریتم Weighted-Majority پیش‌بینی‌ها را بر اساس رأی‌گیری وزن‌داری از استخری از الگوریتم‌های پیش‌بینی[۱۹] انجام می‌دهد و با تغییر این وزن‌ها یاد می‌گیرد. این الگوریتم‌های پیش‌بینی را می‌توان فرضیه‌های متفاوت H در نظر گرفت یا در مقابل می‌تواند از الگوریتم‌های یادگیری مختلفی استفاده کرد. در کل، تنها چیزی که لازم داریم الگوریتمی پیش‌بینی است، که مقدار تابع هدف را برای نمونه پیش‌بینی کند. یکی از خواص جالب الگوریتم Weighted-Majority قابلیت سازگاری آن با داده‌های آموزشی غیر سازگار است. زیرا که این الگوریتم فرضیه‌های ناسازگار با تعدادی نمونه را حذف نمی‌کند و فقط وزن مربوطه را کاهش می‌دهد. خاصیت دوم جالب این الگوریتم است که مرز تعداد خطای این الگوریتم وابسته به مرز تعداد خطای بهترین الگوریتم استخر الگوریتم‌های پیش‌بینی‌اش است.

الگوریتم Weighted-Majority با مقداردهی اولیه‌ی ۱ به تمامی الگوریتم‌های پیش‌بینی شروع می‌شود شروع می‌شود و سپس نمونه‌های آموزشی را دریافت می‌کند. هر بار که الگوریتم پیش‌بینی‌ای نمونه‌ی آموزشی‌ای را اشتباه دسته‌بندی می‌کند وزن مربوطه‌اش با ضریب $0 \leq \beta < 1$ کاهش می‌یابد. تعریف دقیق الگوریتم Weighted-Majority در جدول ۷,۱ آمده است.

---

a_i پیش‌بینی iامین الگوریتم استخر الگوریتم‌های A است. w_i وزن مربوطه به a_i است.

برای تمامی i ها $w_i \leftarrow 1$

برای هر نمونه‌ی آموزشی $< x, c(x) >$

q_0 و q_1 را ۰ مقداردهی اولیه کن

برای هر الگوریتم پیش‌بینی a_i

اگر $a_i(x) = 0$ آنگاه $q_0 \leftarrow q_0 + w_i$

اگر $a_i(x) = 1$ آنگاه $q_1 \leftarrow q_1 + w_i$

اگر $q_1 > q_0$ آنگاه پیش‌بینی کن c(x)=1

اگر $q_0 > q_1$ آنگاه پیش‌بینی کن c(x)=0

اگر $q_0 = q_1$ آنگاه یکی از دو مقدار ۰ یا ۱ را به تصادف برای c(x) پیش‌بینی کن

برای هر الگوریتم پیش‌بینی a_i در A

اگر $a_i(x) \neq c(x)$ آنگاه $w_i \leftarrow \beta w_i$

جدول ۷,۱ الگوریتم Weighted-Majority

[۱۹] pool of prediction algorithms

توجه دارید که اگر β=0 باشد، الگوریتم Weighted-Majority همان الگوریتم Halving خواهد بود. از طرف دیگر اگر مقادیر دیگر β را انتخاب کنیم، دیگر هیچ الگوریتم یادگیری‌ای به طور کامل حذف نخواهد شد. اگر الگوریتمی نمونه‌ی آموزشی‌ای را اشتباه دسته‌بندی کند، خیلی ساده، در رأی با تأثیرگذاری کمتر خواهد داشت.

حال نشان خواهیم داد که مرز تعداد خطای الگوریتم Weighted-Majority را می‌توان با تعداد خطاهای بهترین الگوریتم پیش‌بینی استخر پیش‌بینی‌اش محدود کرد.

قضیه‌ی ۷٫۵. مرز خطای نسبی Weighted-Majority. اگر D سری‌ای از نمونه‌های آموزشی باشد و A نیز مجموعه‌ی n الگوریتم پیش‌بینی باشد و k کمترین تعداد خطای تمامی الگوریتم‌های A برای سری آموزشی D باشد، تعداد اشتباهات الگوریتم Weighted-Majority با β=1/2 حداکثر

$$2.4(k + \log_2 n)$$

خواهد بود.

اثبات. این قضیه را با مقایسه‌ی وزن نهایی بهترین الگوریتم و مجموع وزن‌های دیگر الگوریتم‌ها اثبات می‌کنیم. اگر a_j الگوریتمی از A باشد که مرز خطای بهینه‌ی k را داشته باشد، w_j وزن مربوطه این الگوریتم $\left(\frac{1}{2}\right)^k$ خواهد بود، زیرا که وزن اولیه‌ی برای هر بار اشتباه ضربدر $\frac{1}{2}$ شده است. حال مجموع $W = \sum_{i=1}^{n} w_i$ را که مجموع وزن‌های متناسب n الگوریتم A است را در نظر بگیرید. در ابتدا W مقدار n را داراست. برای هر اشتباه الگوریتم Weighted-Majority این مقدار حداقل به $\frac{3}{4}$ W کاهش می‌یابد. این بدین دلیل است که اکثریت رأی‌گیری وزن‌دار الگوریتم‌ها اشتباه کرده‌اند و ضریب این اکثریت با ضریب $\frac{1}{2}$ کاهش خواهد یافت. اگر M کل تعداد اشتباهات الگوریتم Weighted-Majority برای سری آموزشی D باشد، بنابراین، مجموع کل وزن W حداکثر $\mathrm{n}\left(\frac{3}{4}\right)^{\mathrm{M}}$ خواهد بود. چون وزن نهایی w_j نمی‌تواند بیشتر از وزن کل باشد داریم،

$$\left(\frac{1}{2}\right)^k \leq \mathrm{n}\left(\frac{3}{4}\right)^{\mathrm{M}}$$

با بازنویسی عبارات داریم که،

$$\mathrm{M} \leq \frac{\mathrm{k} + \log_2 n}{-\log_2\left(\frac{3}{4}\right)} \leq 2.4(\mathrm{k} + \log_2 n)$$

و قضیه اثبات می‌شود.

به طور خلاصه، قضیه‌ی بالا نشان می‌دهد که تعداد اشتباهات الگوریتم Weighted-Majority هیچ‌گاه بیشتر از ضریب نسبتی از تعداد اشتباهات بهترین عضو استخر به علاوه‌ی جمله‌ای که رابطه‌ی لگاریتمی با اندازه‌ی استخر دارد نخواهد بود.

این قضیه در حالت کلی توسط (Littlestone and Warmuth 1991) اثبات شده و نشان داده شده که مرز بالا برای مقدار دلخواه $0 \leq \beta < 1$ مرز زیر خواهد بود،

$$\frac{k \log_2\left(\frac{1}{\beta}\right) + \log_2 n}{\log_2 \frac{2}{1+\beta}}$$

۷,٦ خلاصه و منابع برای مطالعه‌ی بیشتر

نکات اصلی این فصل شامل موارد زیر است:

مدل تقریباً درست یا PAC، به الگوریتم‌هایی می‌پردازد که مفاهیم هدف را از کلاس مفاهیم هدف C، با استفاده از نمونه‌های آموزشی تصادفی انتخابی با یک توزیع احتمال ثابت اما نامعلوم یاد می‌گیرد. این مدل یادگیر را ملزم می‌کند که به احتمال حداقل [1-δ] فرضیه‌ای را بیاموزد که تقریباً (با خطای ϵ) درست باشد، با این شرط که پیچیدگی محاسباتی و نمونه‌ای حداکثر به صورت چندجمله‌ای از 1/ϵ، 1/δ، اندازه‌ی مجموعه‌ی نمونه‌ای و اندازه‌ی مفهوم هدف باشد.
با این تعریف از مدل یادگیری PAC، هر یادگیر با فضای فرضیه‌ای متناهی H که C⊆H با احتمال (1-δ) فرضیه‌ای را خروجی خواهد داد که خطای ϵ بر روی مفهوم هدف بعد از m نمونه‌ی آموزشی تصادفی خواهد داشت، به شرط آنکه

$$m \geq \frac{1}{\epsilon}(\ln\left(\frac{1}{\delta}\right) + \ln|H|)$$

این شرط مرزی برای تعداد نمونه‌های کافی برای یادگیر موفق با معیار PAC به ما خواهد داد.

یکی از فرض‌های محدودکننده‌ی مدل PAC این است که یادگیر می‌داند که کلاس مفاهیم C شامل مفهومی که باید یاد گرفته شود است. در مقابل مدل یادگیری agnostic [20] تعریف کلی‌تری می‌کند که یادگیر فرضی در مورد کلاس انتخاب مفهوم هدف ندارد. در مقابل، یادگیر فرضیه‌ای را از H خروجی خواهد داد که کمترین خطا بر روی نمونه‌های آموزشی را داشته باشد (در صورت امکان صفر). تحت این شرایط آزادانه‌تر یادگیری agnostic، یادگیر زمانی اطمینان دارد که با احتمال (1-δ) فرضیه‌ای با خطای ϵ در میان فرضیه‌های H را خروجی می‌دهد که بعد از m نمونه‌ی تصادفی شرط زیر درست باشد:

$$m \geq \frac{1}{2\epsilon^2}(\ln\left(\frac{1}{\delta}\right) + \ln|H|)$$

تعداد نمونه‌های آموزشی لازم برای یادگیری موفق به شدت تحت تأثیر پیچیدگی فضای فرضیه‌ای یادگیر است. یکی از معیارهای مفید پیچیدگی یک فضای فرضیه‌ای H بعد Vapnik-Chervonenkis آن، VC(H) است. VC(H) اندازه‌ی بزرگ‌ترین زیرمجموعه‌ی نمونه‌هاست که می‌توان آن را با H خرد (به تمام روش‌های ممکن تقسیم) کرد.
یک مرز جایگزین تعداد نمونه‌های آموزشی کافی برای یادگیری موفق با مدل PAC با VC(H) بیان می‌شود:

$$m \geq \frac{1}{\epsilon}(4\log_2\frac{2}{\delta} + 8VC(H)\log_2\frac{13}{\epsilon})$$

و یک مرز پایین‌تر نیز:

[20] agnostic learning model

$$m \geq \max\left[\frac{1}{\epsilon} log \frac{1}{\delta}, \frac{VC(H)-1}{32\epsilon}\right]$$

مدل جایگزین دیگری به نام مدل مرز خطا برای بررسی تعداد نمونه‌های آموزشی‌ای که یادگیر قبل از یادگیری کامل مفهوم هدف اشتباه دسته‌بندی می‌کند می‌پردازد. برای مثال الگوریتم Halving حداکثر $\lfloor \log_2 |H| \rfloor$ خطا قبل از یادگیری کامل هر مفهوم هدف از H اشتباه خواهد کرد. برای مفهوم هدف از کلاس C هر الگوریتم در بدترین حالت Opt(C) اشتباه خواهد داشت که:

$$VC(C) \leq Opt(C) \leq \log_2 |C|$$

الگوریتم Weighted-majority از رأی وزن‌دار چندین الگوریتم پیش‌بینی برای دسته‌بندی نمونه‌های جدید استفاده می‌کند. این الگوریتم وزن‌های الگوریتم‌ها را بر اساس تعداد اشتباه در سری‌ای از نمونه‌ها یاد می‌گیرد. جالب است که بدانید که حداکثر تعداد اشتباه این الگوریتم با بهترین الگوریتم استخر رابطه دارد.

اکثر کارهای اولیه تئوری یادگیری محاسباتی با سؤال اینکه آیا یادگیر می‌تواند مفهوم هدف را با داشتن سری‌ای غیر بی‌شمار از داده‌های آموزشی یاد بگیرد سروکار دارند. دسته‌بندی با این محدودیت اولین بار توسط Gold (1967) معرفی شد. تحقیقات کاملی در این زمینه در Angluin (1992) آورده شده است. Vapnik (1982) مفصلاً مسئله‌ی همگرایی یکنواخت بحث کرده و مدل نزدیک به مدل یادگیری PAC را در Vapnik (1984) معرفی می‌کند. بحث ε-exhausting فضای ویژه بر اساس شرح Haussler (1988) پایه‌گذاری شده است. مجموعه‌ی مفیدی از نتایج تحت مدل PAC را می‌توان در Blumer et al. (1989) یافت. Kearns and Vazirani (1994) شرح کاملی از تعدادی زیادی از نتایج تئوری یادگیری محاسباتی را ارائه می‌کند. متون قبلی در این زمینه شامل Anthony and Biggs (1992) و Natarjan (1991) می‌شود.

تحقیقات فعلی بر روی تئوری یادگیری محاسباتی به سمت طیف وسیعی از مدل‌های یادگیری و الگوریتم‌های یادگیری میل می‌کند. بیشتر این تحقیقات را می‌توان در کنفرانس‌های سالانه تئوری یادگیری محاسباتی (COLT) پیدا کرد. تعدادی از مجلات یادگیری ماشین (journal machine learning) نیز به این مبحث اختصاص یافته است.

تمارین

۷,۱ پرسپترونی را با دو ورودی در نظر بگیرید. مرزی برای تعداد نمونه‌های لازم برای اینکه اطمینان داشته باشیم که با احتمال ۹۰٪ حداکثر خطای واقعی ۵٪ را خواهد داشت را بیابید؟ آیا این مرز واقعی به نظر می‌رسد؟

۷,۲ کلاس مفاهیم C را به فرم $(a \leq x \leq b)\wedge(c \leq y \leq d)$ را که در آن a,b,c,d اعداد صحیح درون بازه‌ی (۰،۹۹) هستند را در نظر بگیرید. توجه دارید که این کلاس متناسب با مستطیل‌های حقیقی مقدار در صفحه‌ی xy است. راهنمایی: مربع محصور در بین (0,0) و (n-1,n-1). تعداد مستطیل‌هایی با رئوس اعداد صحیح در این بازه $\left(\frac{n(n+1)}{2}\right)^2$ است.

(a) حد بالای تعداد نمونه‌های تصادفی کافی برای اینکه اطمینان داشته باشیم که برای تمامی مفاهیم هدف c از C، هر یادگیر سازگار با H = C با احتمال ۹۵٪ فرضیه‌ای را خروجی دهد که حداکثر ۰,۱۵ خطا داشته باشد.

(b) حال مستطیلی با مرزهای a, b, c, d را که رئوسش در نقاط حقیقی مقدار است (بجای نقاط صحیح) را در نظر بگیرید. جواب خود را به قسمت اول با این شرایط جدید بدهید.

۷٫۳ در این فصل ما عبارتی برای تعداد نمونه‌ی آموزشی کافی برای اطمینان از اینکه هر فرضیه خطای حقیقی ϵ به اضافه‌ی خطای مشاهده‌ی $error_D(h)$ نداشته باشد پیدا کردیم. در کل، از مرزهای هاپفیلد برای پیدا کردن چنین مرزی استفاده شد (رابطه‌ی ۷٫۳). رابطه‌ی دیگری برای تعداد نمونه‌های آموزشی کافی برای اینکه اطمینان داشته باشیم که هر فرضیه خطای حقیقی کمتر از $(1+\gamma)error_D(h)$ را داشته باشد پیدا کنید. می‌توانید از مرز چرنوف و تعمیم آن برای استخراج چنین نتیجه‌ای استفاده کنید.

مرز چرنوف (chernoff bounds): فرض کنید که $X_1 , \ldots , X_m$، m حاصل مستقل پرتاب سکه (آزمایش برنولی) باشند، با این فرض که احتمال شیر آمدن در هر آزمایش مستقل $\Pr[X_i = 1] = p$ باشد و احتمال خط آمدن $\Pr[X_i = 1] = p - 1$ باشد. اگر $S = X_1 + \cdots + X_m$ مجموع حاصل این پرتاب‌ها باشد، مقدار امید S/m مساوی $E\left[\frac{S}{m}\right] = p$ خواهد بود. مرز چرنوف بر احتمال اینکه این مرز با ضریب $0 \leq \gamma \leq 1$ اختلاف داشته باشد به صورت زیر است:

$$\Pr[S/m > (1+\gamma)p] \leq e^{-mp\gamma^2/3}$$

$$\Pr[S/m < (1-\gamma)p] \leq e^{-mp\gamma^2/2}$$

۷٫۴ مسئله‌ی یادگیری‌ای را در نظر بگیرید که $X = \Re$ مجموعه‌ی اعداد حقیقی و C=H مجموعه‌ی بازه‌های روی اعداد حقیقی به فرم $H = \{(a < x < b) | a, b \in \Re\}$ باشد. احتمال اینکه فرضیه‌ای سازگار با m نمونه از این مفهوم هدف حداقل خطای ϵ داشته باشد چقدر است؟ این سؤال را با استفاده از بعد VC جواب دهید. آیا راه‌حل دیگری بر اساس قوانین اولیه و صرف‌نظر کردن از بعد VC برای جواب به این سؤال وجود دارد؟

۷٫۵ فضای نمونه‌ای X را متناسب با صفحه‌ی x,y در نظر بگیرید. بعد VC فضاهای فرضیه‌ای زیر را مشخص کنید:

(a) مجموعه‌ی تمامی مستطیل‌های صفحه‌ی x,y. $H_r = \{((a < x < b \wedge (c < y < d)) | a, b, c, d \in \Re\}$.

(b) تمامی دایره‌های صفحه‌ی x,y. تمامی نقاط داخل دایره مثبت دسته‌بندی خواهند شد.

(c) مثلث‌های صفحه‌ی x,y. نقاط داخل مثلث مثبت دسته‌بندی خواهند شد.

۷٫۶ یادگیر سازگاری با فضای فرضیه‌ای H_r در مسئله‌ی ۷٫۵ طراحی کنید. مجموعه‌ای از مفاهیم هدف مستطیلی تصادفی متناسب با مستطیل‌های صفحه ایجاد کنید. نمونه‌های تصادفی مربوطه‌ی هر یک از این مفاهیم را با توزیع یکنواخت نمونه‌ها در مستطیل بین (0,0) و (100,100) ایجاد کنید. خطای تعمیم را به عنوان تابعی از تعداد نمونه‌های تصادفی m رسم کنید. در همان شکل رابطه‌ی بین ϵ و m را برای δ=.95 نیز رسم کنید. آیا نتایج با تئوری همخوانی دارد؟

۷٫۷ فضای فرضیه‌ای H_{rd2} را که "درخت‌های تصمیم متوسط با عمق ۲" است را بر روی n متغیر منطقی در نظر بگیرید. درخت‌های تصمیم متوسط با عمق ۲ درخت‌های تصمیمی هستند که (با چهار برگ که تمامی از ریشه فاصله‌ی ۲ دارند) که در آن‌ها بررسی شده در سمت راست و چپ ریشه یکی هستند. برای مثال شکل زیر نمونه‌ای از H_{rd2} است.

```
        x3
       /  \
    x1      x1
   /  \    /  \
  +    -  -    +
```

(a) برحسب n مشخص کنید که اندازه‌ی این مجموعه‌ی H_{rd2} چقدر است؟

(b) مرز بالایی برای تعداد نمونه‌های لازم برای یادگیری با مدل PAC برای یادگیری در H_{rd2} با خطای ϵ و اطمینان δ چقدر است.

(c) الگوریتم Weighted-Majority را برای کلاس H_{rd2} در نظر بگیرید. ابتدا الگوریتم را بـا تمـامی درخت‌هـای درون H_{rd2} بـا وزن اولیه‌ی یکسان ۱ شروع می‌کنیم. هر بار که یک نمونه‌ی جدید مشاهده مـی‌کنیم، آن را بـا اسـتفاده از رأی وزن‌دار تمـامی فرضـیه‌های H_{rd2} دسته‌بندی می‌کنیم. سپس بجای حذف درخت‌های فرضیه‌ای که اشتباه کار می‌کنند فقط وزن تأثیر آن درخت‌ها نصف مـی‌شـود. حـداکثر تعـداد خطاهای را بر حسب n و تعداد خطای بهترین فرضیه‌ی درون H_{rd2} بیان کنید.

۷٫۸ این سؤال به ارتباط بین تحلیل PAC در این فصل و بررسی فرضیه در فصل ۵ می‌پردازد. کار یادگیری‌ای را در نظر بگیرید که نمونـه‌ها بـا n متغیر تصادفی (مثلاً $x_1 \wedge \bar{x}_2 \wedge x_3 \ldots \bar{x}_n$) توصیف می‌شوند و توسط توزیع احتمال ثابت و معلوم $\mathcal{D}$ انتخاب می‌شوند. می‌دانیم کـه مفهـوم هدف عطفی از متغیرهای تصادفی و عکسشان است (مثلاً $x_2 \wedge \bar{x}_5$) و الگوریتم یادگیری از این کلاس مفهوم بـه عنـوان فضـای فرضـیه‌ای H استفاده می‌کند. به یک یادگیر سازگار مجموعه‌ای از ۱۰۰ نمونه انتخابی با $\mathcal{D}$ داده می‌شود. این یادگیر فرضیه‌ی h را از H کـه بـا تمـامی ۱۰۰ نمونه سازگار است خروجی می‌دهد. (بدین معنا که خطای h بر روی این ۱۰۰ نمونه صفر است)

(a) علاقه داریم که خطای واقعی h را که احتمال دسته‌بندی اشتباه نمونه‌های انتخابی با $\mathcal{D}$ است را بیابیم. بر اساس اطلاعات بالا آیا مـی‌توانیـد بازه‌ای را مشخص کنید که خطای واقعی حداقل با احتمال ۹۵٪ در آن قرار گیرد؟ اگر چنین است، بازه را بیان کـرده و بـه وضـوح آن را توجیـه کنید. اگر امکان‌پذیر نیست مشکل را توضیح دهید.

(b) حال مجموعه نمونه‌ی جدید ۱۰۰ تایی دیگری به طور مستقل با همان توزیع $\mathcal{D}$ انتخاب می‌کنید و معلوم می‌شود که h ۳۰ نمونـه از ایـن ۱۰۰ نمونه را اشتباه دسته‌بندی می‌کند. آیا می‌توانید بازه‌ای ارائه کنید که این خطای واقعی با احتمال ۹۵٪ در آن قرار داشته باشـد؟ (بـرای ایـن قسمت از کارایی فرضیه روی نمونه‌های آموزشی صرف‌نظر کنید.) اگر چنین است، بـازه را بیـان کـرده و بـه وضـوح آن را توجیـه کنیـد. اگـر امکان‌پذیر نیست مشکل را توضیح دهید.

(c) ممکن است کمی عجیب به نظر برسد که با این که فرضیه نمونه‌های آموزشی را درست دسـته‌بندی کـرده امـا در دسـته‌بندی نمونـه‌های جدید خطای ۳۰٪ داشته است. احتمال چنین اتفاقی در n های بزرگ بیشتر است یا در n های کوچک‌تر. برای جواب خود توجیه بیاورید.

فرهنگ لغات تخصصی فصل (فارسی به انگلیسی)

ε-exhausted	
آزمایش برنولی	Bernoulli trial
استخری از الگوریتم‌های پیش‌بینی	pool of prediction algorithms
پیچیدگی محاسباتی	Computational complexity
پیچیدگی نمونه‌ای	Sample complexity
ترکیب G	G-composition
توانی	exponential
چارچوب کران خطا	mistake bound framework
چهارچوب	framework
چهارچوب تقریباً درست	probably approximately correct (PAC)
خرد کردن مجموعه‌ای از نمونه‌ها	shattering a set of instances
خطای آموزشی	training error
خطای واقعی	True Error
دقیق	Exact
فرم نرمال فصلی k جمله‌ای	k-term disjunctive normal form (k-term DNF)
قابلیت یادگیری PAC	PAC-learnability
گراف جهت‌دار بدون دور	directed acyclic
گراف لایه‌ای	layered graph
مدل یادگیری تقریباً درست	probably approximately correct (PAC)
مرز بدترین حالت ممکن	wose-case bound
مرز خطا	Mistake bound
مرزهای Hoeffding	Hoeffding bounds
مرزهای اضافی Hoeffding	Hoeffding additive bounds