فصل هفتم : یادگیری محاسباتی

این فصل به صورت تئوری ویژگی های چندین نوع مسئله ی یادگیری ماشین و قابلیت های چندین نوع از الگوریتم های یادگیری ماشین را بیان می کند. این تئوری به دنبال جواب سوالاتی چون "تحت چه شرایطی یادگیری موفق ممکن یا غیرممکن است؟" و "تحت چه شرایطی یک الگوریتم یادگیری خاص موفقیت یادگیری را تظمین می کند؟" است. دو چهارچوب[۱] برای بررسی یادگیری الگورتیم های یادگیری در نظر گرفته می شود. چارچوب اول چهارچوب تقریبا درست[۲] یا (PAC) است، که در آن چارچوب کلاس فرضیه هایی را که می توان یا نمی توان با تعداد چند جمله ای ای از نمونه های آموزشی یادگرفت را بررسی و معیاری طبیعی برای پیچیدگی فضای فرضیه ای که تعداد نمونه های آموزشی برای یادگیری استقرایی را محدود می کند تعریف خواهیم کرد. در چارچوب کران خطا[۳] تعداد خطاهای آموزشی ای را که یادگیر قبل از تعیین فرضیه ی درست انجام می دهد را بررسی خواهیم کرد.

۷.۱ مقدمه

در مطالعه ی یادگیری ماشین این سوال طبیعی است که بپرسیم چه قوانین کلی ای بر یادگیر های ماشین (یا غیر ماشین) حاکم است. آیا می توان کلاس های مسائل یادگیری را که ذاتا سخت یا آسانند را مستقل از الگوریتم یادگیری تعیین کرد؟ آیا می توان تعداد نمونه های لازم برای اینکه یادگیری حتما موفق باشد را تعیین کرد؟ اگر یادگیر بتواند بجای آموزش با دسته ی معینی از نمونه ها آزمایش انجام دهد (در مقابل اینکه نمونه ها به صورت تصادفی به یادگیر داده شوند) این تعداد چگونه تغییر خواهد کرد؟ یا آیا می توان تعداد خطای های یادگیر قبل از یادگیری تابع هدف را مشخص کرد؟ آیا می توان پیچیدگی محاسباتی ذاتی کلاسهای مسائل مختلف را مشخص کرد؟

---

[1] framework
[2] probably approximately correct
[3] mistake bound framework

اگر چه جواب جامع همه ی این سوالات هنوز معلوم نیست، اما قسمت از تئوری هوش محاسباتی برای پاسخ به این سوالات به وجود آمده است. این فصل نتایج کلیدی این تئوری و جواب به این سوالات در بعضی مسائل خاص را در بر می گیرد. در اینجا بحث را به مسئله ی یادگیری استقرایی تابع هدفی نامعلوم از نمونه های آموزشی این تابع هدف و فضای فرضیه ای معلوم محدود می کنیم. با این تعریف مسئله، پاسخ به سوالاتی مثل تعداد نمونه های لازم برای یادگیری موفق و تعداد اشتباهات قبل از یادگیری کامل مطرح می شود. همانطور که بعدا نیز خواهیم دید تعیین مرز های این کمیت ها به ویژگی های مسئله ی یادگیری از جمله موارد زیر وابسته است:

- اندازه یا پیچیدگی فضای فرضیه ای در نظر گرفته شده
- دقت لازم برای یادگیری
- احتمال اینکه یادگیر فرضیه ای موفق را خروجی دهد
- روند ارائه ی نمونه ها

در اکثر موارد، ما بر روی الگوریتم یادگیری خاصی تمرکز نمی کنیم و ترجیح داده می شود بیشتر بر روی کلاسهای الگوریتم های یادگیری با خواص یکسان (فضای فرضیه ای مشابه، نحوه ی نمایش نمونه های آموزشی مشابه و ...) بحث شود. هدف از این فصل پاسخ به سوالاتی نظیر سوالات زیر است:

- پیچیدگی نمونه ای[۱]. تعداد نمونه های آموزشی لازم برای اینکه یادگیر (با احتمال بالایی) به فرضیه ای موفق میل کند؟
- پیچیدگی محاسباتی[۲]. چه میزان محاسبه انجام می شود تا یادگیر با احتمال خوبی به فرضیه ای موفق میل کند؟
- مرز خطا[۳]. تعداد نمونه ها آموزشی ای که یادگیر قبل از همگرا شدن به فرضیه موفق غلت دسته بندی می کند؟

توجه داشته باشید که در بسیاری از حالات چنین سوالاتی مطرح اند. برای مثال، روش های گوناگونی برای تعریف "موفق" وجود دارد. ممکن است یادگیری فرضیه ای را موفق تعریف کنیم که فرضیه ی خروجیش دقیقا مشابه مفهوم هدف باشد. یا در مقابل ممکن است یادگیری را موفق بدانیم که فرضیه اش در اکثر مواقع مشابه مفهوم هدف باشد، یا به طور معمول چنین فرضیه ای را خروجی می دهد. یا به طور مشابه، روند ارائه ی نمونه ها ممکن است متفاوت باشد، ممکن است این نمونه ها توسط یک معلم به یادگیر داده شود یا یادگیر اجازه ی انجام آزمایش داشته باشد یا اینکه نمونه ها توسط یک فرایند تصادفی خارج از کنترل یادگیر انتخاب شوند. همانطور که انتظار می رود، جواب این سوالات به تعریف مسئله و مدل یادگیری وابسته است.

ادامه ی این فصل به صورت زیر ساختار بندی شده است. قسمت ۷.۲ حالت یادگیری احتمالی تقریبا درست (PAC) را معرفی می کند. در ادامه، قسمت ۷.۳ پیچیدگی نمونه ای و پیچیدگی محاسباتی چندین مسئله ی یادگیری را در این حالت بررسی می کند. قسمت ۷.۴ معیار مهمی از پیچیدگی فضا[۴] به نام بعد VC و تاثیر 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 می کند. در انتها نیز، الگوریتم Weighted-Majority را معرفی می کنیم، این الگوریتم روشی برای تلفیق پیشبینی های الگوریتمهای مختلف رقیب است، مرز خطای تئوری این الگوریتم را نیز بررسی خواهیم کرد.

---

[1] Sample complexity
[2] Computational complexity
[3] Mistake bound
[4] space complexity

۷.۲ احتمال یادگیری یک فرضیه ی تقریبا درست

در این بخش حالت خاصی را برای مسائل یادگیری در نظر می گیریم، این حالت مدل یادگیری تقریبا درست (PAC) نامیده می شود. بیایید کار را با این سوال که چه تعداد نمونه ی آموزشی و چه میزان محاسبه لازم است تا کلاس های مختلف یادگیری را با این مدل یاد بگیریم شروع کنیم. برای سادگی کار، بحث را به یادگیری مفاهیم منطقی از داده های آموزشی بدون خطا محدود می کنیم. با این وجود بسیاری از نتایج حاصل را می توان به حالت کلی یادگیری توابع حقیقی مقادیر تابع هدف تعمیم داد (برای مثال به (Natarajan 1991) مراجعه کنید) و بسیاری دیگر از نتایج را می توان به یادگیری از انواع خاصی از داده های خطا دار تعمیم داد (برای مثال به (Laird 1988) و (Kearns and Vazirani 1994) مراجعه کنید).

۷.۲.۱ تعریف مسئله

مشابه فصول گذشته، X مجموعه ی تمامی نمونه های ممکن بر روی تابع هدف مفروض است. برای مثال، X ممکن است مجموعه ی تمامی افراد باشد که با ویژگی های (age (young or old و (height (short or tall باشد. C مجموعه ی مفاهیم هدفی است که ممکن است یادگیر برای یادگیری آنها به کار برده شود. هر مفهوم هدف c در C متناسب با زیر مجموعه ای 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 اسکی باز" باشد. اگر x نمونه ی مثبتی از c باشد، داریم که $c(x)=1$؛ و اگر x نمونه ی منفی ای باشد داریم $c(x)=0$.

در این حالت فرض می کنیم که نمونه ها به صورت تصادفی و با توزیع احتمال $\mathcal{D}$ انتخاب می شوند. برای مثال، $\mathcal{D}$ ممکن است توزیع احتمال نمونه ها افرادی باشد که از یک باشگاه ورزشی در سوئد بیرون می آیند (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 نمی کند. نمونه های آموزشی با این توزیع احتمال انتخاب شده و به همراه مقدار تابع هدف شان $c(x)$ به یادگیر داده می شوند.

یادگیر L مجموعه ای از فرضیه های ممکن مثل H را در یادگیری مفهوم هدف در نظر می گیرد. برای مثال، H ممکن است مجموعه ی تمامی فرضیه های قابل بیان به صورت عطف ویژگی های age و height باشد. بعد از مشاهده ی سری ای از نمونه های آموزشی برای تابع هدف c، L باید فرضیه ای مثل h از H که تخمین آن از c است به عنوان فرضیه ی تخمینی خروجی دهد. موفقیت L را کارایی این فرضیه h بر روی نمونه های جدیدی که به صورت تصادفی از X و با توزیع $\mathcal{D}$ انتخاب می شوند می سنجیم. توزیع احتمال $\mathcal{D}$ همان توزیع احتمالی است که نمونه های آموزشی با انتخاب شده اند.

در چنین حالتی، علاقه ی ما به بررسی کارایی یادگیر های مختلف L با فضای فرضیه های مختلف H در یادگیری مجموعه توابع هدف مختلف درون C است. زیرا که می خواهیم یادگیر L به اندازه ی 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 توزیع های $\mathcal{D}$ را بررسی کنیم.

۷.۲.۲ خطای یک فرضیه

چون علاقه ی ما به نزدیکی فرضیه خروجی یادگیر h به تابع هدف حقیقی c است، بیایید کار را با تعریف خطای واقعی[1] یک فرضیه ی h بر روی c و توزیع احتمال $\mathcal{D}$ شروع کنیم. به صورت غیر رسمی خطای واقعی h، خطای h در دسته بندی نمونه های جدید با توزیع $\mathcal{D}$ است. در واقع این تعریف خطا همان تعریف خطا ی فصل ۵ است. برای راحتی تعریف را برای c که مفهومی منطقی است بازنویسی می کنیم.

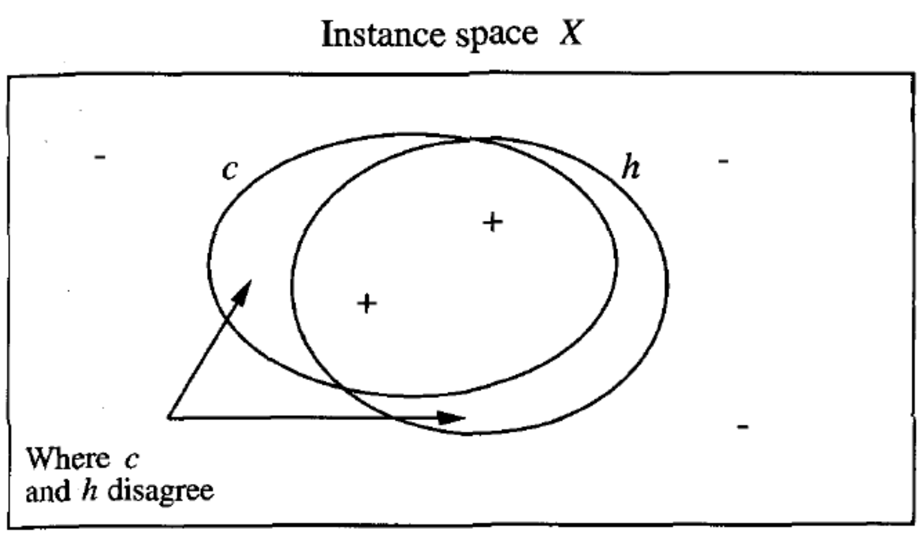
تعریف: خطای واقعی $(error_D(h))$ ی فرضیه ی h برای تابع هدف c و توزیع احتمال نمونه ای $\mathcal{D}$ احتمال این است که نمونه ی انتخابی بر اساس توزیع $\mathcal{D}$ اشتباه دسته بندی شود.

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

در اینجا نماد $\Pr_{x\in D}$ نشان دهنده ی احتمال عبارت با فرض اینکه x از توزیع $\mathcal{D}$ پیروی می کند است.

شکل ۷.۱ این تعریف را به فرم گرافیکی نشان می دهد. مفاهیم c و h با مجموعه ی نمونه های X نمایش داده شده اند، نمونه های آموزشی در این مثال با علامت های + و – نشان داده شده اند. خطای h برای c احتمال دسته بندی 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 کرده ایم که خطای تمامی نمونه های ممکن را اندازه بگیرد و فقط محدود به نمونه های آموزشی نباشد بنابراین انتظار داریم که زمانی که از فرضیه بدست آمده بر روی نمونه های تصادفی جدید استفاده می کنیم چنین خطایی داشته باشند.

توجه دارید که این خطا به شدت به توزیع احتمال نا معلوم $\mathcal{D}$ وابسته است. برای مثال اگر $\mathcal{D}$ توزیعی یکنواخت باشد که به تمامی نمونه های X احتمالی یکسان نسبت می دهد خطای فرضیه ی آمده در شکل ۷.۱ نسبت نمونه های درون ناحیه هلالی به تمامی نمونه ها خواهد بود. با این وجود اگر $\mathcal{D}$ احتمال بیشتری به نمونه های ناحیه هلالی نسبت دهد این خطا بیشتر خواهد شد. و در بد ترین حالت $\mathcal{D}$ احتمال صفر به نمونه های خارج ناحیه هلالی نسبت می دهد و خطا ۱ خواهد بود با وجود اینکه h و c واقعا اشتراک دارند.



[1] True Error

شکل ۷.۱ خطای فرضیه h برای مفهوم هدف c.
خطای h برای مفهوم هدف c احتمال این است که نمونه ای تصادفی در درون ناحیه ای قرار بگیرد که h و c در آن ناحیه دسته بندی های مشابهی ندارنـد. نقاط + و – نشان دهنده ی نمونه های آموزشی مثبت و منفی اند. توجه داشته باشید که با وجود اینکه در تمامی ۵ نمونه ی مشاهده شده دسته بنـدی h و c یکی است، h خطای غیر صفری برای مفهوم هدف c دارد.

بلاخره، توجه داشته باشید که این خطای h برای c به طور مستقیم برای یادگیر غیر قابل مشاهده است. L فقط کارایی h بر روی نمونـه هـای آموزشی را در دسترس دارد و باید انتخاب خود در مورد فرضیه را بر اساس همین معیار انجام دهـد. مـا از عبـارت خطـای آموزشـی[1] (در مقابـل خطای واقعی) برای نمایش نسبت نمونه های آموزشی با دسته بندی اشتباه توسط h به کل نمونه های آموزشی استفاده می کنیم. قسمت بزرگی از بررسی ما از پیچیدگی یادگیری بر محور این سوال متمرکز می شود که "چگونه احتمال دارد که خطای آموزشی مشاهده شـده معیـاری غلـت انداز از $error_{\mathcal{D}}(h)$ باشد؟" است.

به رابطه ی بین این سوال و سوال مطرح شده در فصل ۵ دقت کنید. با توجـه بـه آنچـه در فصـل ۵ گفتـه شـد، خطـای نمونـه ای h را بـرای مجموعه ی S از نمونه ها نسبت دسته بندی اشتباه اعضای S توسط h تعریف شد. خطای آموزشی تعریف شده در بالا خطای نمونه ای S است با این فرض که S مجموعه ی نمونه های آموزشی باشد. در فصل ۵ احتمال اینکه خطای نمونه ای تخمینی غلت انداز از خطای واقعـی باشـد را با این فرض که داده های نمونه ی S مستقل از h باشند بررسی کردیم. اما اینجا حتی این فرض هم درست نیست و فرضیه ی h کاملا 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 ویژگی های توابع هدفی است که می توان آنها را از تعداد معقـولی نمونـه آموزشـی تصـادفی بـا پیچیـدگی محاسـباتی معقـولی یادگرفت.

چه عبارت هایی را می توان درباره ی قابلیت یادگیری یک تابع بیان کرد درست فرض کرد؟ ممکن است سعی کنیم تعداد نمونه هـای آموزشـی لازم برای یادگیری فرضیه ای با $error_{\mathcal{D}}(h) = 0$ را تعیین کنیم. متاسفانه در این تعریف مسئله به دو دلیل این کاری بیهوده اسـت. ابتـدا اینکه برای اینکه به چنین خطایی برسیم باید تمامی نمونه های X را به عنوان نمونه ی آموزش به یادگیر ارائه کنیم (که این فرضی غیر واقعـی است)، و ممکن است چندین فرضیه با مجموعه نمونه های آموزشی سازگار باشند و یادگیر در انتخاب فرضیه ی تخمینی برای مفهوم هدف سـر در گم خواهد ماند. دوم اینکه با معلوم بودن نمونه های آموزشی تصادفی، همیشه احتمالی غیر صفر وجود دارد که نمونه های آموزشـی معیـاری غلت انداز باشد. (برای مثال، با وجود اینکه اغلب قد افراد خارج شده از یک مجموعه ی ورزشی در سوئد متفاوت است اما احتمـال غیـر صـفری وجود دارد که در یک روز تمامی نمونه های مشاهده شده قد ۲ متر داشته باشند).

برای غلبه بر این دو مشکل، شرایط خواستاری مسئله را از دو نظر کاهش می دهیم. ابتدا بجای اینکه شرط کنیم خطای فرضیه صفر شود شرط می کنیم که خطا از مقدار دلخواه کوچک ε کوچکتر باشد. دوم اینکه بجای اینکه شرط کنیم یادگیر روی هر نمونـه ی آموزشـی ممکـن موفـق باشد شرط می کنیم که احتمال عدم موفقیت کمتر از حد دلخواه کوچک خاصی، δ، کمتر باشد. به طور خلاصه شرط می کـنیم کـه یـادگیر بـه صورت احتمالی فرضیه ای تقریبا درست[2] را یاد بگیرد، بنابراین به این فرضیه، فرضیه ی احتمالی تقریبا درست یا PAC می گویند.

[1] training error
[2] approximately correct

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

تعریف: مجموعه ی مفاهیم هدف C را که بر روی نمونه های X با اندازه ی n تعریف شده و یادگیر L با فضای فرضیه ای 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 چند جمله ای از $1/\delta$، $1/\varepsilon$، n و size(c) خروجی باید محاسبه شود.

تعریف ما دو شرط بر روی L می گذارد. ابتدا اینکه L باید با احتمال دلخواه بالایی $(1 - \delta)$ فرضیه ای با مقدار خطای به اندازه ی دلخواه کوچکی ε خروجی دهد. دوم اینکه باید کارایی خوبی داشته باشد، در زمانی حداکثر چندجمله ای از $1/\delta$، $1/\varepsilon$، n و size(c) فرضیه ی خروجی را مشخص کند. در اینجا، n تعداد نمونه های X است. برای مثال، اگر نمونه های X، عطف k ویژگی منطقی باشند، خواهیم داشت که n=k. پارامتر دوم فضا، size(c) است که اندازه ی کدهای c های درون C را نشان می دهد 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 ویژگی ها مشخص شود، size(c) تعداد ویژگی های 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 نمونه های لازم برای یادگیری بیشتر برای ما اهمیت دارد. با این وجود، این دو بسیار به یکدیگر نزدیکند: اگر L نیاز به پردازش حداقلی برای هر نمونه ی آموزشی داشته باشد، برای اینکه C را قابل یادگیری PAC توسط L بدانیم، L حتما به تعداد چند جمله ای ای نمونه ی 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 چند جمله ای ای از نمونه های آموزشی قابل یادگیری است و سپس نشان دهیم زمان محاسبات نیز از چند جمله ای کمتر است.

قبل از رفتن به قسمت بعد، باید به فرض محدود کننده ای در تعریفمان از قابل یادگیری PAC اشاره کنیم. این تعریف مطلقا فرض می کند که فضای فرضیه ای یادگیر H شامل فرضیه هایی است که خطای به اندازه ی دلخواه کوچک برای تمامی مفاهیم درون C دارند. این فرض از این حقیقت ناشی می شود که در تعریف بالا یادگیر زمانی موفق است که بتوان ε را به اندازه ی دلخواه به صفر نزدیک کرد. البته در حالتی که C دقیق معلوم نیست (برای مثال، C در برنامه ای که باید تصاویر چهره را تشخیص دهد چیست؟) تظمین این شرط سخت خواهد بود، مگر اینکه H مجموعه ی 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 نمونه های آموزشی پیدا نمی کند. با این وجود، نتایج حاصل از مدل یادگیری PAC دید مفیدی درباره ی پیچیدگی نسبی مسائل یادگیری مختلف و ضریب بهبود تعمیم با افزایش نمونه های آموزشی به ما می دهد. علاوه بر این، در بخش ۷.۳.۱ این فرض محدود کننده را برای در نظر گرفتن یادگیر بدون پیشفرض حذف می کنیم.

---

[1] PAC learnable

۷.۳ پیچیدگی نمونه ای برای فضای فرضیه ای محدود

همانطور که در بالا نشان دادیم، قابلیت یادگیری PAC [1] به شدت به تعداد نمونه های آموزشی لازم برای یادگیر وابسته است. افزایش تعداد نمونه های آموزشی لازم برای یادگیری متناسب با اندازه ی مسئله، که پیچیدگی نمونه ای مسئله ی یادگیری نامیده می شود، از مهمترین معیار های 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 نمونه های آموزشی است.

در اینجا ما مرزی کلی برای پیچیدگی نمونه ای برای کلاس بزرگی از یادگیر ها، یادگیر های سازگار [2] ارائه می کنیم. یادگیر سازگار یادگیری است که زمانی که ممکن باشد فرضیه ای را که با نمونه های آموزشی سازگار باشد خروجی می دهد. انتظار سازگار بودن یادگیر ها انتظاری دور از ذهن نیست، زیرا که معمولا ما فرضیه ای را که با نمونه های آموزشی سازگار است را به فرضیه های دیگر ترجیح می دهیم. توجه دارید که اکثر الگوریتم های مطرح شده در فصول قبلی از جمله تمامی الگوریتم های فصل ۲ سازگار هستند.

آیا می توانیم مرزی برای تعداد نمونه های آموزشی لازم برای هر یادگیر سازگار که مستقل از الگوریتم است پیدا کنیم؟ جواب این سوال بلی است. برای ایجاد چنین مرزی بد نیست که بعضی تعاریف فصل ۲ را درباره ی فضای ویژه بازگو کنیم. در آنجا فضای ویژه ی $VS_{H,D}$ را مجموعه ی تمامی فرضیه های $h \in H$ را که نمونه های آموزشی D را درست دسته بندی می کنند 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 فرضیه ای از فضای ویژه را خروجی می دهد. دلیل این نتیجه در تعریف فضای ویژه مشهود است، زیرا که فضای ویژه تمامی فرضیه های سازگار در H با نمونه های آموزشی را در بر می گیرد. بنابراین برای محدود کردن تعداد نمونه های آموزش لازم برای هر یادگیر سازگار کافی است تعداد نمونه های آموزشی لازم را برای تظمین اینکه فضای ویژه فرضیه ای غیر قابل قبول را در بر نگیرد معلوم کنیم. تعریف زیر، که به نام Haussler (1988) نام گذاری شده، این شرط را به صورت دقیق مشخص می کند.

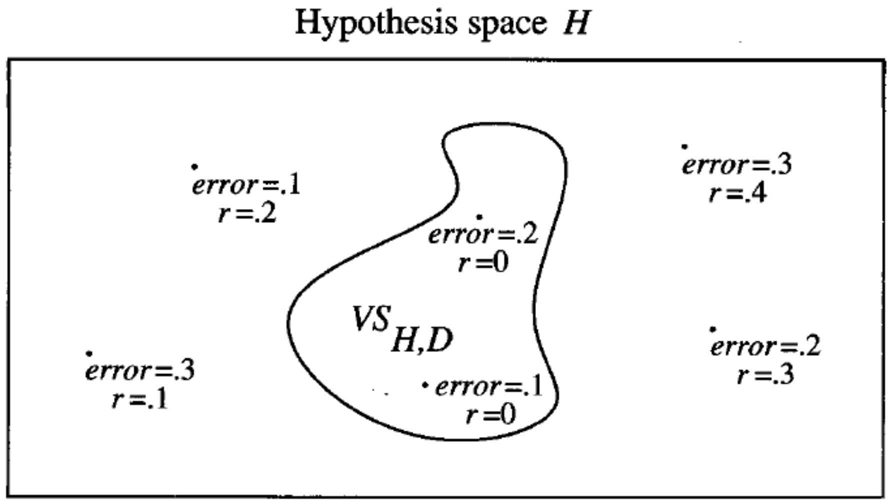
تعریف: فضای فرضیه ای H، مفهوم هدف c، توزیع نمونه ای $\mathcal{D}$ و مجموعه ی نمونه های آموزشی D که برای آموزشی c است را در نظر بگیرید. فضای ویژه ی $VS_{H,D}$ زمانی ε-exhausted برای c و D است که برای هر فرضیه ی h در $VS_{H,D}$ خطایی کمتر از ε برای c و D داشته باشیم.

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

تعریف بالا در شکل ۷.۲ نمایش داده شده است. فضای ویژه زمانی ε-exhausted است که تمامی فرضیه های سازگار با نمونه های آموزشی مشاهده شده (برای مثال، آنهایی که خطای نمونه ای صفر دارند) خطایی کمتر از ε داشته باشند. البته از دید یادگیر فقط فرضیه هایی که به طور کامل با نمونه های آموزشی سازگار اند قابل تشخیص است، همگی آنها خطای آموزشی صفر خواهند داشت. فقط شاهدی که از ماهیت مفهوم هدف اگاه است می تواند با قطعیت فضای ویژه ی ε-exhausted را مشخص کند. جالب است که بررسی ای احتمالی به ما اجازه می دهد

[1] PAC-learnability
[2] consistent learners

که احتمال اینکه فضای ویژه بعد از تعدادی نمونه ی آموزشی ε-exhausted باشد را بدون اینکه اطلاعاتی در مورد ماهیت مفهوم هدف یا توزیع نمونه های آموزشی داشته باشیم محدود کنیم. Haussler (1988) چنین مرزی را با قضیه ی زیر ایجاد می کند.



شکل ۷.۲ exausted کردن فضای ویژه.
فضای ویژه ی $VS_{H,D}$ زیر مجموعه ای از فرضیه های h∈H است که خطای آموزشی صفر دارند (در شکل با r=0 نشان داده شده است). البته خطای واقعی $error_{\mathcal{D}}(h)$ است (که در شکل با error نمایش داده شده) که حتی ممکن است برای فرضیه های فضای ویژه غیر صفر باشد. فضای ویژه زمانی ε-exausted است که تمامی فرضیه های باقیمانده ی درون $VS_{H,D}$ داشته باشیم $error_{\mathcal{D}}(h) < \varepsilon$.

قضیه ی ۷.۱ فضای ویژه ی ε-exhausted. اگر فضای فرضیه ای H محدود باشد و D نیز سرای ای از m≥1 نمونه ی تصادفی مستقل از مفهوم هدف c باشد، برای هر 0≤ε≤1 احتمال اینکه فضای ویژه ی $VS_{H,D}$ 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 فرضیه های درون H باشند که خطای واقعی بیشتر از ε برای c دارند. اگر و فقط اگر حداقل یکی از این k فرضیه با تمامی نمونه های آموزشی سازگار باشد فضای ویژه ε-exhausted نخواهد بود. احتمال اینکه فرضیه ای که خطای واقعی بیشتر از ε دارد با نمونه ای که به صورت اتفاقی انتخاب می شود سازگار باشد حداکثر (1-ε) است. بنابراین احتمال اینکه این فرضیه با m نمونه ی مستقل سازگار باشد $(1-\varepsilon)^m$ خواهد بود. حال اگر k فرضیه خطایی بیشتر از ε داشته باشند، احتمال اینکه حداقل یکی از این فرضیه ها با تمامی m نمونه ی 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. بلاخره، از رابطه ی 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 می رسد.

این قضیه کران بالایی بر حسب تعداد نمونه های آموزشی m و حداکثر خطای مجاز ε واندازه ی H برای احتمال اینکه فضای ویژه -ε exhausted نباشد ارائه می کند. اما از نظر دیگر، این مرز احتمال اینکه m نمونه ی آموزشی در حذف تمامی فرضیه های "بد" (فرضیه هایی که خطای واقعی بیشتر از ε دارند) در یادگیر سازگار با فضای فرضیه ای H موفق نشوند را نشان می دهد.

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

$$|\mathrm{H}|\mathrm{e}^{-\varepsilon m} \leq \delta \tag{7.1}$$

با بازنویسی رابطه برای m داریم که

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

به طور خلاصه نامساوی رابطه ی ۷.۲ مرزی کلی برای تعداد نمونه های آموزشی لازم برای اینکه تمامی یادگیر های سازگار با موفقیت هر مفهوم هدف درون H را برای مقادیر دلخواه δ و ε یادبگیرند را مشخص می کند. این تعداد نمونه ی آموزشی برای تظمین اینکه هر فرضیه ی سازگار تقریبا (با احتمال (1-δ) درست) درست (حداکثر با خطای ε) باشد را تعیین می کند. توجه دارید که m به صورت خطی با 1/ε و لگاریتمی 1/δ متناسب است. همچنین با نسبت لگاریتمی با اندازه ی فضای فرضیه ای H نیز متناسب است.

توجه دارید که مرز بالا می تواند ذاتا اغراق آمیز باشد. برای مثال، با وجود اینکه احتمال اینکه فضای ویژه باید در بازه ی [0,1] قرار بگیرد، اما این مرز با افزایش |H| به صورت خطی افزایش پیدا می کند. برای فضای های فرضیه ای به اندازه ی کافی بزرگ، این مرز می تواند به راحتی بزرگتر از یک شود. نتیجه اینکه مرز نامساوی ۷.۲ می تواند ذاتا برای تعداد نمونه های آموزشی اغراق آمیز باشد. ضعف این مرز معمولا در جمله ی |H| است که از اثبات هنگام جمع احتمالات یک فرضیه ی غیر قابل قبول در میان تمامی فرضیه ها ایجاد می شود. در واقع، در بسیاری مواقع مرزی کوچکتر فضاهای فرضیه ای بینهایت بزرگ را محدود می کند. این مرز موضوع قسمت ۷.۴ خواهد بود.

۷.۳.۱ یادگیری agnostic و فرضیه های غیر سازگار

رابطه ی ۷.۲ از این جهت اهمیت دارد که تعداد نمونه های آموزشی لازم برای تظمین اینکه (با احتمال (1-δ)) هر فرضیه ی H که خطای آموزشی صفر دارد خطای واقعی حداکثر ε داشته باشد را تعیین می کند. متاسفانه اگر H شامل تابع هدف c نباشد، همیشه نمی توان فرضیه ای پیدا کرد که خطای آموزشی صفر داشته باشد. در چنین حالتی، از یادگیر می خواهیم که فرضیه ای را خروجی دهد که کمترین خطای ممکن را بر روی نمونه های آموزشی داشته باشد. یادگیری که هیچ پیشفرضی در مورد تابع هدف نمی کند و فقط فرضیه ای از H را که کمترین خطای آموزشی دارد را خروجی می دهد، یادگیری agnostic نامیده می شود، زیرا که هیچ فرض قبلی ای برای اینکه آیا C⊆H درست است یا خیر نمی کند.

با وجود اینکه رابطه ی ۷.۲ بر این فرض که یادگیر فرضیه ای با خطای صفر را خروجی می دهد پایه گذاری شده است، اما مرزی مشابه را می توان برای حالت کلی تری که یادگیر فرضیه ای با خطای آموزشی غیر صفر را خروجی می دهد می توان بدست اورد. به عبارت دقیقتر فرض کنید که D مجموعه ی خاصی از نمونه ای آموزشی موجود است (البته با $\mathcal{D}$ که توزیع نمونه ای است متفاوت است) و $error_D(h)$ خطای نمونه ای فرضیه ی h بر روی D است. در کل، $error_D(h)$ نسبت اشتباه های h بر روی نمونه های 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 نمونه ای است متفاوت است. حال فرض کنید

h_{best} نماد فرضیه ای از H باشد که کمترین خطای نمونه ای را بر روی نمونه های آموزشی دارد. چه تعداد نمونه ی آموزش لازم است تا تظمین شود که (با احتمال قوی) خطای واقعی $error_{\mathcal{D}}(h_{best})$ کمتر یا مساوی $error_{D}(h_{best})$ باشد؟ توجه دارید که رابطه ی این سوال با سوال مطرح شده در قسمت قبلی این است که سوال قسمت قبلی حالت خاصی از این سوال بود (حالتی که $error_{D}(h_{best}) = 0$).

این سوال را می توان با استفاده از تشابه با اثبات قضیه ی ۷.۱ جواب داد (به رابطه ی ۷.۳ مراجعه کنید). یادآوری مرز های Hoeffding [1] (که گاهی مرز های اضافی Hoeffding [2] نامیده می شود.) در اینجا مفید است. مرز های Hoeffding مشتق بین احتمال واقعی یک اتفاق و میزان مشاهده ی آن اتفاق در m آزمایش مستقل را بررسی می کند. به عبارت دقیقتر، این مرز ها به m آزمایش مستقل برنولی[3] اعمال می شوند (برای مثال در m پرتاب سکه ای با احتمال شیر آمدنی خاص). این کاملا مشابه تعریف مسئله ی ما در تخمین خطای فرضیه در فصل ۵ است: احتمال اینکه شیر بیاید مشابه احتمال این است که فرضیه یک نمونه ی تصادفی را اشتباه دسته بندی کند. m پرتاب مستقل سکه مشابه انتخاب m نمونه ی مستقل از توزیع نمونه ای است. نسبت تعداد شیر ها به کل m پرتاب مشابه نسبت دسته بندی های اشتباه به کل m نمونه ی تصادفی است.

مرز های Hoeffding می گوید که اگر خطای نمونه ای $error_{D}(h)$ بر روی مجموعه ی D شامل m نمونه ی تصادفی باشد، خواهیم داشت که:

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

این رابطه مرزی برای احتمال اینکه یک فرضیه ی دلخواه خطای نمونه ای بسیار گمراه کننده داشته باشد را به ما می دهد. برای اینکه مطمئن باشیم که بهترین فرضیه پیدا شده توسط L حداکثر خطایی با این مرز دارد، باید احتمال اینکه هر فرضیه از |H| فرضیه های موجود خطای بزرگی داشته باشند را در نظر گرفت.

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

اگر این احتمال را δ بنامیم و به دنبال تعداد نمونه های لازم برای کمتر بودن δ از مقدار خاصی بگردیم به این رابطه می رسیم که:

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

این رابطه تعمیم رابطه ی ۷.۲ برای حالتی است که یادگیر هنوز بهترین فرضیه h∈H را انتخاب می کند و خطای نمونه ای بهترین فرضیه نیز می تواند غیر صفر باشد. توجه دارید که m با H و 1/δ رابطه ای لگاریتمی دارد و همانطور که در مشاهده می شود که به حالت خاص تر ۷.۲ می رسیم. با این وجود در این حالت کلی تر m به جای رابطه ی خطی متناسب با مجذور 1/ε است.

[1] Hoeffding bounds
[2] Hoeffding additive bounds
[3] Bernoulli trial

۷.۳.۲ عطف عبارات منطقی PAC-Learnable است

حال که مرزی برای تعیین تعداد نمونه های آموزشی کافی برای اینکه بتوان با احتمال خوبی تابع هدف را یادگرفت بدست آورده ایم، از این مرز می توانیم برای تعیین پیچیدگی نمونه ای و Pac-learnable بودن دسته ی خاصی از مفاهیم هدف استفاده کرد.

مجموعه ی مفاهیم هدف C را که با عطف عبارات منطقی بیان می شود را در نظر بگیرید. یک عبارت منطقی می تواند هر متغیر منطقی (مثل Old) یا نقیضش (مثل Old¬) باشد. بنابراین عطف عبارات منطقی مثل توابع هدفی چون "Old ∧ ¬Tall" را نیز شامل می شود. آیا C قابل یادگیری PAC است؟ می توان نشان داد که پاسخ چنین سوالی آری است. کافی است ابتدا نشان دهیم هر یادگیر سازگار فقط تعداد چند جمله ای ای نمونه ی آموزشی برای یادگیری هر c در C لازم دارد و الگوریتمی ارائه کنیم که زمانی چند جمله ای 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 فرضیه ای H که مشابه C است استفاده می کند. از رابطه ی ۷.۲ می توان برای محاسبه ی تعداد m نمونه ی آموزشی تصادفی کافی تا یادگیر L با احتمال (1-δ) فرضیه ای خروجی با ماکزیمم خطای ε بدهد استفاد کرد. برای این کار، لازم است که |H| را که اندازه ی فضای فرضیه ای مربوطه است را تعیین کرد.

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

با اضافه کردن $|H| = 3^n$ در رابطه ی ۷.۲ مرز پیچیدگی نمونه ای یادگیری عطف n عبارت منطقی به صورت زیر بدست می آید.

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

برای مثال اگر یک یادگیر سازگار، یادگیری بخواهد تابع هدفی توصیفی با ۱۰ عبارت و با احتمال درستی بیش از ۹۵ درصد فرضیه ای با خطای کمتر از .1 را یاد بگیرد، برای m که تعداد نمونه های 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 رابطه ی خطی و مستقیم با n (تعداد عبارات فضای فرضیه ای)، 1/ε و رابطه ای لگاریمتی با 1/δ دارد. اما رابطه ی 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 چند جمله ای برای هر نمونه داشته باشد و کل محاسبات نیز کمتر از چند جمله ای کل نمونه های آموزشی باشد، مسلما محاسبات کل نیز کمتر از چند جمله ای تعداد نمونه ها خواهد بود.

در یادگیری عطف عبارات منطقی، یکی از الگوریتم هایی که شرایط لازم را دارد در فصل ۲ مورد بحث قرار گرفت. این الگورتیم Find-S است، که خاص ترین فرضیه ی سازگار با نمونه های آموزشی را محاسبه می کند. برای هر نمونه ی مثبت آموزشی جدید، الگوریتم اشتراک بین عباراتی فرضیه ی فعلی و نمونه ی آموزشی جدید را در زمانی با رابطه ی خطی با n محاسبه می کند. بنابراین، الگوریتم Find-S کلاس مفاهیم عطفی n عبارت منطقی و نقیضشان را به فرم PAC یاد می گیرد.

قضیه ی ۷.۲ قابلیت یادگیری PAC عطف عبارات منطقی. کلاس C که مجموعه ی عطف عبارات منطقی است توسط الگوریتم Find-S با استفاده از H=CFind-S قابل یادگیری PAC است.

اثبات. رابطه ی ۷.۴ نشان می دهد پیچیدگی نمونه ای این کلاس مفاهیم نسبت به $1/\delta$ و $1/\varepsilon$ چندجمله ای و از size(c) مستقل است. برای پردازش مرحله به مرحله ی هر نمونه ی 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، رابطه ی ۷.۲ پایه ای کلی برای محدود کردن پیچیدگی یادگیری توابع مفهوم کلاس معلوم C ارائه می کند. در بالا این رابطه را برای کلاس عطف عبارات منطقی به کار بردیم. به طور مشابه می توان نشان داد که بسیاری از کلاسهای مفهوم پیچیدگی نمونه ای چند جمله ای دارند. (تمرین ۷.۲)

۷.۳.۳.۱ یادگیر های بدون بایاس

همه ی کلاسهای مفهوم مرز پیچیدگی نمونه ای با رابطه ی ۷.۲ محدودی ندارند. برای مثال کلاس مفاهیم بایاس نشده ی C را که تمامی مفاهیم قابل تعلیم بر روی X را در بر می گیرد را در نظر بگیرید. مجموعه ی C تمامی مفاهیم هدف قابل تعریف همان مجموعه ی توانی X، مجموعه ی تمامی زیرمجموعه های X، خواهد بود که $|C| = 2^{|X|}$. فرض کنید که نمونه های درون X با n متغیر منطقی تعریف شوند، بنابراین خواهیم داشت که، $|X| = 2^n$ بنابرای خواهیم داشت که $|C| = 2^{2^n}$. البته برای یادگیری چنین کلاس بایاس نشده ی یادگیری، خود یادگیر نیز باید از فضای فرضیه ای بدون بایاس استفاده کند H=C. با جایگزاری $|H| = 2^{2^n}$ در رابطه ی ۷.۲ پیچیدگی نمونه ای برای یادگیر مفاهیم بدون بایاس روی X مشخص می شود.

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

بنابراین، این کلاس بدون بایاس از مفاهیم هدف بنابر رابطه ی ۷.۲ پیچیدگی نمونه ای نمایی (exponential) در مدل PAC دارد. با وجود اینکه رابطه ی ۷.۲ پیچیدگی نمونه ای با اغراق برای کلاس مفاهیم بدون بایاس را توانی از n می داند اما در حقیقت اثبات می شود که این مرز اغراق آمیز نیست.

۷.۳.۳.۲ مفاهیم k-term DNF و k-CNF

همچنین می توان کلاس های مفاهیمی را پیدا کرد که پیچیدگی نمونه ای چند جمله ای دارند اما با این وجود نمی توان آنها را در زمانی چند جمله ای یادگرفت. یکی از مثالهای جالب چنین کلاسهایی، کلاس مفاهیم فرم نرمال فصلی k جمله ای (k-term DNF) [1] 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 نقیض هایشان 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}

[1] k-term disjunctive normal form

تخمین بالایی از H است زیرا که در شرایطی که $T_i = T_j$ و T_i کلی تر از T_j است را دوبار می شماریم. با این وجود می توان از مرز حداکثری |H| برای پیدا کردن مرز حداکثری پیچیدگی نمونه ای استفاده کرد، با جایگذاری در رابطه ی ۷.۲ خواهیم داشت،

$$m \geq \frac{1}{\varepsilon}\left(nk\ln 3 + \ln\left(\frac{1}{\delta}\right)\right) \qquad (7.6)$$

این رابطه نشان می دهد که پیچیدگی نمونه ای k-term DNF چند جمله ای ای از 1/ε، 1/δ، n و k است. با این وجود که پیچیدگی نمونه ای از درجه چند جمله ای است، پیچیدگی محاسباتی از درجه چند جمله ای نیست، زیرا می توان نشان داد که این مسئله یادگیری معادل دیگر مسائل یادگیری است که در زمان چند جمله ای قابل حل نیستند (مگر اینکه RP=NP). بنابراین، با وجود اینکه k-term DNF پیچیدگی چند جمله ای دارد، اما پیچیدگی محاسباتی آن برای یادگیری که در آن H=C از درجه ی چند جمله ای 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 چند جمله ای از نمونه ها دارند و پیچیدگی نمونه ای چند جمله ای دارد. این کلاس بزرگتر کلاس نمایش های 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 زیر مجموعه ی k-CNF است بسیار ساده است زیرا که می توان هر عبارت k-DNF را به سادگی با یک عبارت k-CNF بازنویسی کرد (برعکس این قضیه بر قرار نیست). با این وجود که k-CNF نسبت به k-DNF شامل تر است، هم پیچیدگی نمونه ای چند جمله ای و هم پیچیدگی محاسباتی چند جمله ای دارد. بنابراین، کلاس مفاهیم k-DNF با الگوریتم یادگیری ای که از H=k-CNF استفاده می کند قابل یادگیری PAC است. برای بحث دقیقتر به (Kearns and Vazirani 1994) مراجعه کنید.

۷.٤ پیچیدگی نمونه ای برای فضاهای فرضیه ای بیکران

در بخش بالا نشان دادیم که پیچیدگی نمونه ای برای یادگیری PAC متناسب با لگاریتم اندازه ی فضای فرضیه ای است. با وجود اینکه رابطه ی ۷.۲ رابطه ی بسیار مفیدی است اما دو اشکال در بیان پیچیدگی نمونه ای بر اساس |H| وجود دارد. ابتدا اینکه ممکن است به مرز های ضعیفی ختم گردد (مقدار δ می تواند برای مقادیر بزرگ |H| به شکل قابل توجهی بزرگتر از یک باشد). دوم اینکه در فضای فرضیه ای بیکران به طور کلی نمی توان از رابطه ی ۷.۲ استفاده کرد.

در اینجا معیار دیگری از پیچیدگی H به نام بعد H Vapnik-Chervonenkis (به اختصار بعد H VC یا VC(H)) را معرفی خواهیم کرد. همانطور که در ادامه نیز خواهیم دید، می توان مرز پیچیدگی را با معیار VC(H) به جای |H| بیان کرد. در بسیاری از موارد، مرز پیچیدگی بر پایه ی VC(H) قوی تر از مرز رابطه ی ۷.۲ خواهد بود. به علاوه این مرز امکان بررسی پیچیدگی نمونه ای بسیاری از فضاهای فرضیه ای بیکران را نیز فراهم می کند.

۷.٤.۱ خرد کردن مجموعه ای از نمونه ها

بعد VC پیچیدگی فضای فرضیه ای H را نه بر اساس |H| و بلکه بر اساس تعداد نمونه های متمایز X که می توانند به کلـی بـا H مشـخص شوند بیان می کند.

برای دقیقتر کردن این نمادگذاری، بیایید ابتدا نمادگذاری خرد کردن مجموعه ای از نمونه ها[1] را مشخص کنیم. زیـر مجموعـه ای از نمونـه هـا مانند $S \subseteq X$ را در نظر بگیرید. برای مثال، شکل ۷.۳ زیر مجموعه ای از X شامل سه نمونه را نشان می دهد. هر فرضیه ی h در H S را بـه دو مجموعه تقسیم می کند؛ این دو مجموعه، مجموعه های $\{x \in S|h(x) = 0\}$ و $\{x \in S|h(x) = 0\}$ هسـتند. بـا معلـوم بـودن مجموعه ی S می توان $2^{|S|}$ تقسیم دوتایی مختلفی که اعضای H ممکن بعضی از آنها را نتوانند ایجاد کنند نوشت. زمانی می گـوییم H، S را خرد می کند که برای هر یک از تقسیم های دوتایی S را بتوان با فرضیه ای از H نمایش داد.

تعریف. مجموعه ی نمونه های S با فضای فرضیه ای H خرد می شود اگر و فقط اگر برای هر تقسیم دوتایی S فرضیه ای سازگار وجود داشته باشد.

شکل ۷.۳ مجموعه ای از سه نمونه را نشان می دهد که توسط فضای فرضیه ای خرد می شود. توجه دارید کـه بـرای هـر یـک از 2^3 تقسـیم دوتایی این سه نمونه فرضیه ای وجود دارد.

توجه دارید که اگر مجموعه ای از نمونه ها توسط یک فضای فرضیه ای خرد نشود، بدین معناست که فرضیه ای (تقسـیم دوتـایی ای) بـر روی نمونه ها وجود دارد که نمی توان آنرا با فضای فرضیه ای نشان داد. قدرت خرد کردن یک فضای فرضـیه ای بـرای مجموعـه ای از نمونـه هـا معیاری از قدرت نمایش این فضای فرضیه ای برای نمایش مفاهیم تعریف شده بر روی این مجموعه از نمونه هاست.

۷.٤.۲ بعد Vapnik-Chervonenkis

قدرت خرد کردن مجموعه ای از نمونه ها رابطه ی نزدیکی با بایاس استقرایی یک فضای فرضیه ای دارد. با توجه به آنچه در فصل ۲ گفته شد، فضای فرضیه ای بدون بایاس فضای فرضیه ای است که می تواند تمامی مفاهیم (تقسیم های دوتایی) قابـل تعریـف روی فضـای نمونـه ی را نمایش بدهد. اما اگر H نتواند X را خرد کند، اما در مقابل بتواند زیر مجموعه ی بزرگی از X را خرد کند چه؟ به نظر مـی رسـد اینکـه هـر قـدر زیرمجموعه ی خردشده ی X بزرگتر باشد، H نیز شامل تر خواهد بود. بعد VC ی H به طور دقیقتر معیار زیر است.

تعریف. بعد Vapnik-Chervonenkis، VC(H)، برای فضای فرضیه ای H کـه بـر روی فضـای نمونـه ای X تعریـف شـده انـدازه ی بزرگترین زیرمجموعه ی کراندار X است که با H خرد می شـود. اگـر H بتوانـد هـر زیرمجموعـه ی دلخـواه X را خـرد کنـد خـواهیم داشـت، VC(H)≡∞.

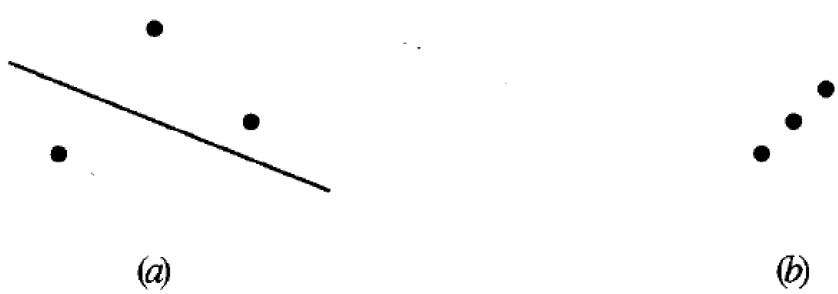
توجه دارید که برای تمامی H های کراندار داریم $VC(H) \leq \log_2 |H|$. برای درک این رابطه فرض کنید داریـم $VC(H) = d$. پـس H به 2^d فرضیه ی محض برای خرد کردن d نمونه احتیاج خواهد داشت، پس $2^d \leq |H|$ و $d = VC(H) \leq \log_2 |H|$.

[1] shattering a set of instances

۷.٤.۲.۱ چندین مثال

برای پیدا کردن درکی از VC(H)، چند نمونه فضای فرضیه ای را در نظر بگیرید. برای شروع، فرض کنیـد کـه X مجموعـه ی اعـداد حقیقـی $X=\mathfrak{R}$است (که قد افراد را توصیف می کند) و H مجموعه ی بازه های اعداد حقیقی است. به عبارت دیگر H مجموعه ی فرضیه هـایی اسـت که به فرم a<x<b بیان می شوند و a و b نیز اعداد ثابت حقیقی اند. VC(H) در اینجا چیست؟ برای جـواب بـه ایـن سـوال، بایـد بزرگتـرین زیرمجموعه ی X را پیدا کنیم که با H خرد شود. زیر مجموعه ی دو عضوی خاصی از اعداد حقیقی مثل S={3.1,5.7} را در نظر بگیرید. آیـا می توان S را با H خرد کرد؟ بله، برای مثال چهار فرضیه ی (1<x<2)، (1<x<4)، (4<x<7) و (1<x<7) این کار را انجام می دهند. این چهار مجموعه با هم هر یک از تقسیمهای دو عضوی S را، مثل دربر داشتن هیچکدام، یکی و هر دو نمونه، را نمایش می دهند. از آنجـایی کـه مجموعه ای دو عضوی پیدا کردیم که با H خرد شود پس VC(H) حداقل دو خواهد بود. اما آیا مجموعه ای با اندازه ی سه وجـود دارد کـه H آنرا خرد کند؟ مجموعه ی $S=\{x_0,x_1,x_2\}$ با سه نمونـه ی دلخـواه را در نظـر بگیریـد. بـدون از دسـت دادن کلیـت فـرض کنـیم کـه $x_0<x_1<x_2$. واضح است که این زیرمجموعه را نمی توان خرد کرد، زیرا که تقسیم دو عضوی ای که x_0 و x_2 را در برگرفتـه و x_1 را در بر نگیرد را نمی توان با یک فرضیه نشان داد. بنابراین، هیچ زیرمجموعه ی سه عضوی را نمی توان 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 صفحه ی x-y باشد (شکل ۷.۴). فرض کنیم H تمام سطوح تصمیم گیری خطـی بـر روی ایـن صـفحه است. به عبارت دیگر، H فضای فرضیه ای متناسب با یک واحد پرسپترون با دو ورودی است (برای بحث کلی تر به فصل ۴ مراجعه کنید). بعـد VC ی H چیست؟ درک اینکه هر دو نقطه در فضا را می توان با فضای فرضیه ای H خرد کرد بسیار آسـان اسـت، زیـرا کـه چهـار خـط پیـدا می شود که هیچکدام، یکی و یا هردو نمونه را در بر بگیرند؟ اما در مورد مجموعه های سه نقطه ای چه؟ تا زمانی که نقاط بـر روی یـک خـط نیستند، 2^3 خط پیدا خواهیم کرد که مجموعه را خرد کند. البته مجموعه ی سه نقطه ی روی یک خط را نمی توان خرد کرد (به همان دلیل که نمی توانستیم در مثال قبل مجموعه ی سه نقطه ای از روی خط حقیقی را خرد کنیم). در چنین شرایطی VC(H) چند است، دو یا سه؟ حـداقل سه است. تعریف بعد VC می گوید که اگر بتوانیم مجموعه ای d عضوی را پیدا کنیم که بتوان آنرا با H خرد کرد داریم کـه $VC(H)\geq d$. برای نشان دادن اینکه $VC(H)<d$ باید نشان دهیم که مجموعه ی d عضوی را نمی توان با H خرد کرد. در این مثال، هیچ مجموعـه ی چهار عضوی را نمی توان خرد کرد پس VC(H)=3. به طور کلی می توان نشان داد که بعد VC ی سطوح تصمیم خطـی در فضـای r بعـدی (بعد VC پرسپترونی با r ورودی) r+1 است.



شکل ۷.۴ بعد VC ی سطوح تصمیم خطی در صفحه ی x-y ۳ است.

(a) مجموعه ای از سه نقطه که با سطوح تصمیم خطی خرد می شود. (b) مجموعه ای از سه نقطه که نمی توان آنرا 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 فرضیه ی H عطف حداکثر سه عبارت منطقی نمونه ها باشد. VC(H) چیست؟ می توان نشان داد که این مقدار حداقل ۳ است. هر یک از نمونه ها را با رشته ای سه بیتی متناسب با عبارات l_1، l_2 و l_3 نشان می دهیم. مجموعه ی سه نمونه ای زیر را در نظر بگیرید:

$instance_1: 100$

$instance_2: 010$

$instance_3: 001$

این مجموعه ی سه عضوی از نمونه ها را می توان با H خرد کرد، زیرا که برای هر تقسیم دوتایی می توان به فرم زیر فرضیه ای ساخت: اگر فرضیه ای نمونه ی $instance_i$ را در بر نمی گیرد $\neg l_i$ را به فرضیه اضافه کن. برای مثال فرضیه ای را در نظر بگیرید که $instance_2$ را دربرگرفته اما $instance_1$ و $instance_3$ را دربر نمی گیرد. از فرضیه ی $\neg l_1 \wedge \neg l_2$ برای این حالت استفاده خواهیم کرد. این بحث را می توان به سادگی از ۳ ویژگی به n ویژگی تعمیم داد. در واقع، VC(H) در این حالت دقیقا n است، اما نشان دادن این کار کمی سخت تر است زیرا باید نشان دهیم که مجموعه ی n+1 عضوی ای وجود ندارد که با H خرد شود.

۷.٤.۳ پیچیدگی نمونه ای و بعد VC

در قسمتهای قبلی سوال "چه تعداد نمونه ی تصادفی برای تخمین یکی از فرضیه های C کافی است؟" (چه تعداد نمونه ی 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 می توان جوابی دیگر برای این سوال پیدا کرد، مشابه آنچه پیشتر با مرز رابطه ی ۷.۲ بیان کردیم. این مرز جدید به فرم زیر است (به (Blumer et al. 1989) مراجعه کنید)

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

توجه دارید که مشابه رابطه ی ۷.۲ تعداد نمونه های لازم m با لگاریتم 1/δ متناسب است. اما به جای رابطه ی خطی با لگاریتم برابر رابطه خطی با 1/ε متناسب است. قابل توجه است که، جمله ی ln|H| که در مرز قبلی بود با معیار جایگزین پیچیدگی فضای فرضیه ای، VC(H)، جایگزین شده است. (توجه دارید که $VC(H) \leq \log_2|H|$).

رابطه ی ۷.۷ کران بالایی برای تعداد نمونه های آموزشی لازم برای اینکه هر فرضیه ی C را به طور PAC با ε و δ دلخواه یاد بگیریم را معلوم می کند. پیدا کردن کران پایین برای این تعداد با استفاده از قضیه ی زیر امکان پذیر است (به (Ehrenfeucht et al. 1989) مراجعه کنید).

قضیه ی ۷.۳ کران پایین پیچیدگی نمونه ای. 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 فرضیه ای را خروجی می دهد که $error_{\mathcal{D}}(h) > \varepsilon$.

این قضیه نشان می دهد که اگر تعداد نمونه های آموزشی بسیار کم باشد، هیچ یادگیری نمی تواند تمامی مفاهیم هدف C غیربدیهی را به طور PAC یاد بگیرد. بنابراین، این قضیه کران پایینی بر روی تعداد نمونه های آموزشی برای یاگیری موفق را ارائه می کند، این مرز کامل کننده ی کران بالای ذکر شده برای تعداد کافی نمونه های آموزشی است. توجه دارید که این کران پایین با پیچیدگی کلاس مفاهیم C بیان می شود، در حالی که کران بالا با H تعیین می شد. (چرا؟)

این کران پایین نشان می دهد که کران بالای نامساوی ۷.۷ به اندازه ی کافی محکم است. هر دو کران با $1/\delta$ رابطه ی لگاریتمی و با VC(H) رابطه ی خطی دارند. تنها تفاوت های باقی مانده در این دو کران وابستگی کران بالا به $\log(1/\varepsilon)$ است.

۷.٤.٤ بعد VC برای شبکه های عصبی

با درنظر داشتن بحث شبکه های عصبی مصنوعی از فصل ۴، تعیین بعد VC شبکه ای از واحدهای مرتبط، مثل شبکه های عصبی تک سویه که توسط فرایند backpropagation آموزش داده می شوند، جالب خواهد بود. این بخش نتیجه ی کلی ای از محاسبه ی بعد VC شبکه های بدون دور را برا اساس ساختار شبکه و بعد VC خود واحد ها را ارائه می کند. این بعد VC را می توان برای محدود کردن نمونه های آموزشی لازم برای یادگیری تقریبا درست شبکه ی تک سویه برای مقادیر دلخواه ε و δ به کاربرد. می توانید در اولین مطالعه ی کتاب این بخش را بدون از دست دادن پیوستگی مطلب نخوانید.

شبکه ی G متشکل از واحدها با گراف بدون دور را در نظر بگیرید. یک گراف جهت دار[۱] بدون دور[۲] گرافی است که یالهایش جهت دارند (واحد ها ورودی و خروجی هستند) و دور ندارد. گراف لایه ای[۳] گرافی است که گره هایش را بتوان به صورتی تقسیم بندی کرد که تمامی یالهای جهت دار خروجی از گره های لایه l به گره های لایه ی l+1 بروند. گراف شبکه ی عصبی تک سویه در فصل ۴، مثالی از چنین گرافهای جهت دار لایه ای بدون دور است.

ثابت می شود که می توان بعد VC چنین شبکه هایی را بر اساس ساختارشان و بعد VC واحد های اولیه ی سازندیشان محدود کرد. برای فرموله کردن این حقیقت، باید ابتدا چندین عبارت دیگر را تعریف کنیم. بیایید فرض کنیم که n تعداد ورودی های شبکه ی G است و این شبکه تنها یک خروجی دارد. فرض کنیم که واحد های داخلی G، N_i (هر واحدی که ورودی نباشد) حداکثر r ورودی داشته و از تابعی منطقی مقدار $c_i: \mathfrak{R}^r \to \{0,1\}$ از کلاس توابع C استفاده کند. برای مثال اگر واحد های داخلی پرسپترون باشند C کلاس توابع خطی مقدار آستانه ای تعریف شده بر روی $\mathfrak{R}^r$ خواهد بود.

حال می توانیم ترکیب G[۴] ی C را به عنوان کلاس تمام توابعی که شبکه G می تواند به کار ببرد با این فرض که هر واحد شبکه ی G یکی از توابع کلاس C را مورد استفاده قرار دهد تعریف کرد. به طور خلاصه، ترکیب G ی C فضای فرضیه ای است که توسط شبکه ی G قابل نمایش است.

---

[1] directed
[2] acyclic
[3] layered graph
[4] G-composition

قضیه ی زیر بعد VC ترکیب G ی C را بر اساس بعد VC ی C و ساختار G محدود می کند.

قضیه ی ۷.۴. بعد VC شبکه های جهت دار لایه ای بدون دور. (برای اطلاعات بیشتر به (Kearns and Vazirani 1994) مراجعه کنید). فرض کنید G گراف جهت دار لایه ای 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 پایه ی لگاریتم طبیعی (عدد نپر) است.

توجه دارید که این مرز بعد VC شبکه ی G رابطه ی خطی با بعد VC ی d تک واحد هایش و رابطه ی لگاریتمی با s، تعداد واحدهای آستانه ی شبکه دارد.

شبکه های لایه ای بدون دوری را در نظر بگیرید که از گره های پرسپترون تشکیل یافته اند. 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 می کند. 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 می توان به همراه قضیه ی بالا برای محدود کردن بعد VC شبکه ای لایه ای 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 می توان تعداد m نمونه ی آموزشی کافی برای یادگیری (با احتمال حداقل (1-δ)) هر مفهوم هدف $\mathrm{C}_{\mathrm{G}}^{\mathrm{perceprtons}}$ را با خطای ε مشخص کرد. با جایگزاری رابطه ی بالا برای VC شبکه در رابطه ی ۷.۷ خواهیم داشت،

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

همانطور که با این مثال شبکه ی پرسپترن نشاند داده شد، قضیه ی بالا از این نظر جالب است که متدی کلی برای محدود کردن بعد VC ی شبکه ای بدون دور و لایه ای از واحد ها را بر اساس ساختار شبکه و بعد VC تک واحد سازنده محدود می کنیم. متاسفانه نتایج بالا مستقیما در مورد شبکه هایی که با Backpropagation آموزش داده می شوند صادق نیست، به دو دلیل. اول اینکه این نتایج برای شبکه های پرسپترون، و نه واحدهای سیگموید که در backpropagation مورد استفاده است، نتیجه گیری شده است. با این وجود، توجه دارید که بعد VC ی واحدهای سیگموید حداقل به اندازه ی بعد VC ی واحدهای پرسپترون است، زیرا که واحد سیگموید می تواند پرسپترون را تا حد دلخواه با افزایش وزنها تخمین بزند. بنابراین، مرز بالا برای m حداقل مرز ممکن برای شبکه های بدون دور واحدهای سیگموید است. مشکل دوم تعمیم نتیجه ی بالا این است که Backpropagation از شبکه ای با وزنهای غیر صفر کار خود را شروع کرده و با تغییر وزنها به فرضیه ی قابل قبول می رسد. بنابراین، Backpropagation با معیار توقف cross-validation بایاسی استقرایی با ترجیح شبکه هایی با وزنهای کوچکتر دارد. این بایاس استقرایی که به طور موثری VC را کاهش می دهد 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 مسئله های دیگر و دیگر سوالات را در نظر دربر می گیرد. تعریف مسئله های یادگیری مختلفی که مورد مطالعه قرار گرفته است در نحوه ی ایجاد نمونه های یادگیری (مشاهده ی سوم شخص[1] نمونه های تصادفی، انتخاب آزمایش توسط یادگیر)، نویز داده ها (با خطا یا بدون خطا)، تعریف موفق (مفهوم هدف باید دقیقا یادگرفته شود یا اینکه تقریبا یا با احتمال خاصی یادگرفته شود)، فرضهای یادگیر (شامل توزیع نمونه ای و اینکه $C \subseteq H$) و معیاری که با آن یادگیر ارزیابی می شود (تعداد نمونه های آموزشی، تعداد اشتباه ها، زمان کل یادگیری) متفاوت اند.

در این بخش به مدل یادگیری مرز خطا، که در آن یادگیر با تعداد اشتباه هایش قبل از همگرایی به فرضیه ی درست ارزیابی می شود خواهیم پرداخت. مشابه تعریف مسئله ی PAC، فرض می کنیم یادگیر سری ای از نمونه های آموزشی را دریافت می کند. با این وجود، در اینجا می خواهیم یادگیر قبل از دریافت هر نمونه ی x مقدار تابع هدف c(x) را (قبل از معلوم شدن مقدار درست هدف توسط آموزش دهنده) پیشبینی کند. سوال مطرح این است که "یادگیر قبل از یادگیری مفهوم هدف چه تعداد پیشبینی اشتباه خواهد کرد؟" اهمیت این سوال در کاربرد عملی است، زیرا که یادگیری باید زمانی که سیستم درحال استفاده واقعی است انجام شود، نه در مرحله ی آموزشی مجزا. برای مثال، اگر سیستم برای یادگیری پیشبینی اینکه چه پرداخت های credit cart ی باید ثبت شود و چه پرداخت هایی تقلبی هستند بر اساس اطلاعاتی که حین استفاده از سیستم جمع آوری می کند طراحی می شود، بنابراین علاقه خواهیم داشته که تعداد اشتباهات قبل از همگرایی به تابع هدف مینیمم شود. در اینجا تعداد کل اشتباهات می تواند اهمیت بیشتری نسبت به تعداد کل نمونه های آموزشی داشته باشد.

این مسئله ی یادگیری مرز خطا را می توان 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 می گیریم. یادگیری مفهوم هدف به طور دقیق بدین معناست که به فرضیه ای میل کنیم که $(\forall x) h(x) = c(x)$.

۷.۵.۱ مرز خطا ی الگوریتم Find-S

برای تصور، دوباره فضای فرضیه ای H عطف n عبارت منطقی $l_1 \ldots l_n$ و نقایضشان را در نظر بگیرید (برای مثال Rich∧¬Handsome). الگورتیم Find-S که خاصترین فرضیه سازگار با نمونه های آموزشی را محاسبه می کرد را از فصل ۲ به یاد بیاورید. یکی از سرراست ترین پیاده سازی الگوریتم Find-S برای فضای فرضیه ای H در زیر آمده:

Find-S:

فرضیه ی h را با خاص ترین فرضیه $l_1 \wedge \neg l_1 \wedge l_2 \wedge \neg l_2 \ldots l_n \wedge \neg l_n$ مقدار دهی اولیه کن.

برای هر نمونه ی مثبت x

هر عبارت h را که توسط x راضی نمی شد را حذف کن.

فرضیه ی h را خروجی بده.

[1] passive

Find-S به طور حدی به فرضیه ای میل می کند که خطایی نخواهد داشت، به شرطی که C⊆H باشد و داده های آموزشی نیز بدون خطا باشند. Find-S با خاصترین فرضیه (فرضیه ای که تمامی نمونه ها را منفی دسته بندی می کند) شروع می کند، سپس به صورت پلکانی این فرضیه را در مواقع لازم برای پوشاندن نمونه های آموزشی مثبت کلی تر می کند. برای نمایش فرضیه ای استفاده شده در اینجا، این کلی سازی با حذف عبارات راضی نشده خواهد بود.

آیا می توان 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 نمونه ی منفی ای را مثبت دسته بندی نخواهد کرد. دلیل این است که فرضیه ی فعلی h همیشه حداقل به اندازه ی مفهوم هدف c خاص است. بنابراین، برای محاسبه ی تعداد اشتباهات، فقط باید اشتباهاتی را که نمونه ی مثبت، منفی دسته بندی می شود را بشماریم. این گونه اشتباهات قبل از یادگیری کامل c چند بار اتفاق خواهند افتاد؟ اولین نمونه ی مثبت ارائه شده به Find-S را در نظر بگیرید. یادگیر در دسته بندی این نمونه دقیقا یک اشتباه انجام خواهد داد، زیرا که فرضیه ی اولیه ی یادگیر تمامی نمونه ها را منفی دسته بندی می کند. با این وجود، نتیجه این خواهد بود که نصف 2n عبارت فرضیه ی اولیه حذف خواهد شد و n عبارت باقی خواهد ماند. برای هر نمونه ی مثبت بعدی که به اشتباه منفی دسته بندی می شود حداقل یکی از این n عبارت از فرضیه ی 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 کلی ترین مفهوم هدف $(\forall x)c(x)=1$ را یاد بگیرد بدترین سری نمونه ها، یعنی نمونه هایی که در هر بار اشتباه فقط یک عبارت را حذف می کنند 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 می گیرد، و پلکانی در این فضای ویژه با برخورد با نمونه های جدید تجدید نظر می کند. الگوریتم های Candidate-Elimination و List-Then-Eliminate در فصل ۲ چنین الگوریتم هایی هستند. در این بخش مرز بدترین حالت ممکن[1] را روی تعداد اشتباهاهایی که چنین یادگیری انجام می دهد را برای فضای فرضیه ای محدود H، دوباره با فرض اینکه تابع هدف را باید دقیقا یادبگیریم، محاسبه می کنیم.

برای بررسی تعداد اشتباهات حین یادگیری، ابتدا بادید تعیین کنیم که یادگیر دقیقا چگونه دسته بندی نمونه ی جدید را پیشبینی می کند. بیایید فرض کنیم که این پیشبینی با رای گیری در بین فرضیه های فضای ویژه فعلی انجام می گیرد. اگر اکثر فرضیه های فضای ویژه نمونه ی جدید را مثبت دسته بندی 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 رای اکثریت برای پیشبینی های بعدی، گاهی الگوریتم Halving نامیده می شود. ماکزیمم تعداد اشتباهات الگوریتم Halving برای H محدود دلخواه قبل از یادگیری مفهوم هدف چیست؟ توجه دارید که یادگیری "دقیق"[2] مفهوم هدف، متناسب با رسیدن به حالتی است که فضای ویژه فقط شامل یک فرضیه شود. (مشابه معمول، فرض می کنیم که مفهوم هدف c در H وجود دارد).

[1] wose-case bound
[2] exact

برای بدست آوردن مرز خطا، توجه داشته باشید که الگوریتم Halving فقط زمانی اشتباه می کند که اکثریت فضای ویژه اش نمونه ی جدید را اشتباه دسته بندی کنند. در چنین شرایطی، هنگامی که دسته بندی جدید برای یادگیر آشکار می شود، اندازه ی فضای ویژه حداقل به نصف اندازه ی فعلی اش کاهش می یابد (فقط فرضیه هایی که اقلیت بودند باقی می مانند). با معلوم بودن اینکه با هر اشتباه اندازه ی فضای ویژه حداقل به نصف کاهش می یابد و با دانستن اینکه فضای ویژه ی اولیه فقط |H| عضو داشته، حداکثر اشتباهات ممکن قبل از اینکه فضای ویژه فقط یک عضو داشته باشد $\log_2 |H|$ است. در واقع می توان نشان داد که مقدار این مرز $\lfloor \log_2 |H| \rfloor$ است. برای مثال، فرض کنید که |H|=7 است. اولین اشتباه اندازه ی |H| را به ۳ و دومین اشتباه اندازه ی آنرا 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 یادبگیرد. دلیل این حقیقت این است که حتی زمانی که رای اکثریت درست است، الگوریتم فرضیه های 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 رای فرضیه هاست. فصل ۶ دسته بندی کننده ی بهینه ی بیز، که از رای گیری وزن دار میان فرضیه ها استفاده می کند را معرفی می کند. در دسته بندی کننده ی بهینه ی بیز، وزن نسبت داده شده به هر فرضیه احتمال ثانویه ی تخمینی توصیف مفهوم هدف به شرط مشاهده ی داده های آموزشی است. در ادامه ی این بخش الگوریتمی متفاوت بر پایه ی رای گیری وزن دار دیگری را به نام Weighted-Majority معرفی خواهیم کرد.

۷.۵.۳ مرز های خطای بهینه

بررسی بالا مرز بدترین حالت اشتباه را برای دو الگوریتم خاص، Find-S و Candidate-Elimination، تعیین کرد. تعیین اینکه مرز بهینه ی 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 الگوریتم های یادگیری ممکن است. به عبارت دقیقتر، برای یادگیری الگوریتم A و مفهوم هدف c $M_A(c)$ را ماکزیمم خطای A در تمامی سری های ممکن نمونه های آموزشی برای یادگیری دقیق c تعریف می کنیم. حال برای هر کلاس فرضیه ای غیر تهی تعریف می کنیم که $M_A(C) \equiv \max_{c \in C} M_A(c)$. توجه دارید که دربالا 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 $M_{Halving}(C) \leq \log_2(|C|)$ برای تمامی کلاسهای مفهوم C است.

مزر خطای بهینه را برای کلاس مفاهیم C به فرم زیر تعریف می کنیم.

تعریف. اگر C کلاس غیرتهی دلخواهی باشد، مرز بهینه ی خطای C، که با Opt(C) نمایش داده می شود، مینیمم روی تمام الگوریتم های ممکن A $M_A(C)$ است.

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

به طور غیر رسمی، این تعریف Opt(C) را تعداد اشتباهات سخت ترین C با استفاه از سخت ترین سری نمونه های آموزشی برای بهترین الگوریتم یادگیری تعریف می کند. (Littlestone 1987) نشان می دهد که برای هر کلاس مفهوم C، رابطه ای جالب میان مرز خطای بهینه ی C و مرز خطای الگوریتم Halving و بعد VC ی C وجود دارد،

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

علاوه براین کلاسهای مفاهیمی وجود دارد که این چهار کمیت برای آنها دقیقا مساوی است. یکی از چنین کلاس های مفاهیم کلاس مجموعه ی توانی $\mathrm{C_p}$ برای مجموعه ی محدود X است. در چنین شرایطی $\mathrm{VC}(\mathrm{C_p}) = |X| = \log_2(|\mathrm{C_p}|)$، بنابراین تمامی چهار کمیت بالا برابرند. (Littlestone 1987) نمونه هایی از دیگر کلاسهای مفاهیمی که VC(C) مطلقا کمتر از Opt(C) و Opt(C) مطلقا از $\mathrm{M_{Halving}(C)}$ کمتر است ارائه می کند.

۷.۵.٤ الگوریتم Weighted-Majority

در این قسمت تعمیمی از الگوریتم Halving را به نام الگوریتم Weighted-Majority بررسی می کنیم. الگوریتم Weighted-Majority پیشبینی ها را بر اساس رای گیری وزن داری از استخری از الگوریتم های پیشبینی[1] انجام می دهد و با تغییر این وزن ها یاد می گیرد. این الگوریتم های پیشبینی را می توان فرضیه های متفاوت H در نظر گرفت یا در مقابل می تواند از الگوریتم های 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 داده های آموزشی غیر سازگار است. زیرا که این الگوریتم فرضیه های ناسازگار با تعدادی نمونه را حذف نمی کند و فقط وزن مربوطه را کاهش می دهد. خاصیت دوم جالب این الگوریتم است که مرز تعداد خطای این الگوریتم وابسته به مرز تعداد خطای بهترین الگوریتم استخر الگوریتم های پیشبینی اش است.

الگورتیم Weighted-Majority با مقدار دهی اولیه ی ۱ به تمامی الگوریتم های پیشبینی شروع می شود شروع می شود و سپس نمونه های آموزشی را دریافت می کند. هر بار که الگوریتم پیشبینی ای نمونه ی آموزشی ای را اشتباه دسته بندی می کند وزن مربوطه اش با ضریب $0 \leq \beta < 1$ کاهش می یابد. تعریف دقیق الگوریتم Weighted-Majority در جدول ۷.۱ آمده است.

---

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

برای تمامی i ها $w_i \leftarrow 1$

برای هر نمونه ی آموزشی $< x, c(x) >$

q_0 و q_1 را ۰ مقدار دهی 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 $a_i(x) \neq c(x)$ انگاه $w_i \leftarrow \beta w_i$

---

[1] pool of prediction algorithms

جدول ۷.۱ الگورتیم Weighted-Majority

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 الگورتیم یادگیری ای به طور کامل حذف نخواهد شد. اگر الگوریتمی نمونه ی آموزشی ای را اشتباه دسته بندی کند، خیلی ساده، در رای با تاثیرگذاری کمتر خواهد داشت.

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

قضیه ی ۷.۵. مرز خطای نسبی Weighted-Majority. اگر D سری ای از نمونه های آموزشی باشد و A نیز مجموعه ی n الگوریتم پیشبینی باشد و k کمترین تعداد خطای تمامی الگوریتم های 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 مقایسه ی وزن نهایی بهترین الگوریتم و مجموع وزنهای دیگر الگوریتم ها اثبات می کنیم. اگر a_j الگوریتمی از A باشد که مرزخطای بهینه ی k را داشته باشد، w_j وزن مربوطه این الگوریتم $\left(\frac{1}{2}\right)^k$ خواهد بود، زیرا که وزن اولیه ی 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 می یابد. این بدین دلیل است که اکثریت رای گیری وزن دار الگوریتم ها اشتباه کرده اند 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 نمی تواند 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 می شود.

به طور خلاصه، قضیه ی بالا نشان می دهد که تعداد اشتباهات الگوریتم Weighted-Majority هیچگاه بیشتر از ضریب نسبتی از تعداد اشتباهات بهترین عضو استخر به علاوه ی جمله ای که رابطه ی لگاریتمی با اندازه ی 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 $0 \le \beta < 1$ 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 مطالعه ی بیشتر

نکات اصلی این فصل شامل موارد زیر می باشد:

مدل تقریبا درست یا PAC، به الگوریتم هایی می پردازد که مفاهیم هدف را از کلاس مفاهیم هدف C، با استفاده از نمونه های آموزشی تصادفی انتخابی با یک توزیع احتمال ثابت اما نامعلوم یاد می گیرد. این مدل یادگیر را ملزم می کند که به احتمال حداقل $[1-\delta]$ فرضیه ای را بیاموزد که تقریبا (با خطای ϵ) درست باشد، با این شرط که پیچیدگی محاسباتی و نمونه ای حداکثر به صورت چند جمله ای ای از $1/\epsilon$، $1/\delta$، اندازه ی مجموعه ی نمونه ای و اندازه ی مفهوم هدف باشد.

با این تعریف از مدل یادگیری PAC، هر یادگیر با فضای فرضیه ای متناهی H که $C \subseteq H$ با احتمال $(1-\delta)$ فرضیه ای را خروجی خواهد داد که خطای ϵ بر روی مفهوم هدف بعد از m نمونه ی آموزشی تصادفی خواهد داشت، به شرط آنکه

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

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

یکی از فرضهای محدود کننده ی مدل PAC این است که یادگیر می داند که کلاس مفاهیم C شامل مفهومی که باید یاد گرفته شود می باشد. در مقابل مدل یادگیری agnostic [1] تعریف کلی تری می کند که یادگیر فرضی در مورد کلاس انتخاب مفهوم هدف ندارد. در مقابل، یادگیر فرضیه ای را از H خروجی خواهد داد که کمترین خطا بر روی نمونه های آموزشی را داشته باشد (در صورت امکان صفر). تحت این شرایط آزادانه تر یادگیری agnostic، یادگیر زمانی اطمینان دارد که با احتمال $(1-\delta)$ فرضیه ای با خطای ϵ در میان فرضیه های H را خروجی می دهد که بعد از m نمونه ی تصادفی شرط زیر درست باشد:

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

تعداد نمونه های آموزشی لازم برای یادگیری موفق به شدت تحت تاثیر پیچیدگی فضای فرضیه ای یادگیر است. یکی از معیار های مفید پیچیدگی یک فضای فرضیه ای H بعد Vapnik-Chervonenkis آن، VC(H) است. VC(H) اندازه ی بزرگترین زیر مجموعه ی نمونه هاست که می توان انرا با H خرد (به تمام روش های ممکن تقسیم) کرد.

یک مرز جایگزین تعداد نمونه های آموزشی کافی برای یادگیری موفق با مدل PAC با VC(H) بیان می شود:

---

[1] agnostic learning model

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

و یک مرز پایین تر نیز:

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

مدل جایگزین دیگری به نام مدل مرز خطا برای بررسی تعداد نمونه های آموزشی ای که یادگیر قبل از یادگیری کامل مفهوم هدف اشتباه دسته بندی می کند می پردازد. 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 بد ترین حالت Opt(C) اشتباه خواهد داشت که:

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

الگوریتم Weighted-majority از رای وزندار چندین الگوریتم پیشبینی برای دسته بندی نمونه های جدید استفاده می کند. این الگوریتم وزنهای الگوریتم ها را بر اساس تعداد اشتباه در سری ای از نمونه ها یاد می گیرد. جالب است که بدانید که حداکثر تعداد اشتباه این الگوریتم با بهترین الگوریتم استخر رابطه دارد.

اکثر کارهای اولیه تئوری یادگیری محاسباتی با سوال اینکه آیا یادگیر می تواند مفهوم هدف را با داشتن سری ای غیر بیشامار از داده های آموزشی یاد بگیرد سر و کار دارند. دسته بندی با این محدودیت اولین بار توسط Gold (1967) معرفی شد. تحقیقات کاملی در این ضمینه در Angluin (1992) اورده شده است. Vapnik (1982) مفصلا مسئله ی همگرایی یکنواخت بحث کرده و مدل نزدیک به مدل یادگیری PAC را در Vapnik (1984) معرفی می کند. بحث ε-exhausting فضای ویژه بر اساس شرح Haussler (1988) پایه گذاری شده است. مجموعه ی مفیدی از نتایج تحت مدل PAC را می توان در Blumer et al. (1989) یافت. Kearns and Vazirani (1994) شرح کاملی از تعدادی زیادی از نتایج تئوری یادگیری محاسباتی را ارائه می کند. متون قبلی در این ضمینه شامل Anthony and Biggs (1992) و Natarjan (1991) می شود.

تحقیقات فعلی بر روی تئوری یادگیری محاسباتی به سمت طیف وسیعی از مدلهای یادگیری و الگوریتم های یادگیری میل می کند. بیشتر این تحقیقات را می توان در کنفرانس های 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 نمونه های 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 می رسد؟

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 بازه ی (۰،۹۹) هستند را در نظر بگیرید. توجه دارید که این کلاس متناسب با مستطیل های حقیقی مقدار در صفحه ی xy است. راهنمایی: مربع محصور در بین (0,0) و (n-1,n-1). تعداد مستطیل هایی با رئوس اعداد صحیح در این بازه $\left(\frac{n(n+1)}{2}\right)^2$ است.

(a) حد بالای تعداد نمونه های تصادفی کافی برای اینکه اطمینان داشته باشیم که برای تمامی مفاهیم هدف c از C، هر یادگیر سازگار با H = C با احتمال ۹۵٪ فرضیه ای 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 نمونه ی آموزشی کافی برای اطمینان از اینکه هر فرضیه خطای حقیقی ϵ به اضافه ی خطای مشاهده ی $error_D(h)$ نداشته باشد پیدا کردیم. در کل، از مرزهای هاپفیلد برای پیدا کردن چنین مرزی استفاده شد (رابطهی ی ۷.۳). رابطه ی دیگری برای تعداد نمونه های آموزشی کافی برای اینکه اطمینان داشته باشیم که هرف فرضیه خطای حقیقی کمتر از $(1+\gamma)error_D(h)$ را داشته باشد پیدا کنید. می توانید از مرز چرنوف و تعمیم آن برای استخراج چنین نتیجه ای 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 $0 \le \gamma \le 1$ اختلاف داشته باشد به صورت زیر است:

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

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

۷.۴ مسئله ی یادگیری ای را در نظر بگیرید که $X = \Re$ مجموعه ی اعداد حقیقی و C=H مجموعه ی بازه های روی اعداد حقیقی به فرم $H = \{(a < x < b) | a, b \in \Re\}$ باشد. احتمال اینکه فرضیه ای سازگار با m نمونه از این مفهوم هدف حداقل خطای ϵ داشته باشد چقدر است؟ این سوال را با استفاده از بعد VC جواب دهید. آیا راه حل دیگری بر اساس قوانین اولیه و صرف نظر کردن از بعد VC برای جواب به این سوال وجود دارد؟

۷.۵ فضای نمونه ای X را متناسب با صفحه ی x,y در نظر بگیرید. بعد VC فضاهای فرضیه ای زیر را مشخص کنید:

(a) مجموعه ی تمامی مستطیل های صفحه ی x,y. $H_r = \{((a < x < b \wedge (c < y < d)) | a, b, c, d \in \Re\}$.

(b) تمامی دایره های صفحه ی x,y. تمامی نقاط داخل دایره مثبت دسته بندی خواهند شد.

(c) مثلثهای صفحه ی x,y. نقاط داخل مثلث مثبت دسته بندی خواهند شد.

۷.۶ یادگیر سازگاری با فضای فرضیه ای H_r در مسئله ی ۷.۵ طراحی کنید. مجموعه ای از مفاهیم هدف مستطیلی تصادفی متناسب با مستطیلهای صفحه ایجاد کنید. نمونه های تصادفی مربوطه ی هر یک از این مفاهیم را با توزیع یکنواخت نمونه ها در مستطیل بین (0,0) و (100,100) ایجاد کنید. خطای تعمیم را به عنوان تابعی از تعداد نمونه های تصادفی m رسم کنید. در همان شکل رابطه ی بین ϵ و m را برای δ=.95 نیز رسم کنید. آیا نتایج با تئوری همخوانی دارد؟

۷.۷ فضای فرضیه ای 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 فاصله ی ۲ دارند) که در آنها بررسی شده در سمت راست و چپ ریشه یکی هستند. برای مثال شکل زیر نمونه ای از H_{rd2} است.

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

(a) برحسب n مشخص کنید که اندازه ی این مجموعه ی H_{rd2} چقدر است؟

(b) مرز بالایی برای تعداد نمونه های 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 با تمامی درختها ی درون H_{rd2} با وزن اولیه ی یکسان ۱ شروع می کنیم. هر بار که یک نمونه ی جدید مشاهده می کنیم، آنرا با استفاده از رای وزن دار تمامی فرضیه های H_{rd2} دسته بندی می کنیم. سپس بجای حذف درختهای فرضیه ای که اشتباه کار می کنند فقط وزن تاثیر آن درخت ها نصف می شود. حداکثر تعداد خطاهای را بر حسب n و تعداد خطای بهترین فرضیه ی درون H_{rd2} بیان کنید.

۷.۸ این سوال به ارتباط بین تحلیل PAC در این فصل و بررسی فرضیه در فصل ۵ می پردازد. کار یادگیری ای را در نظر بگیرید که نمونه ها با n متغیر تصادفی (مثلا $x_1 \wedge \bar{x}_2 \wedge x_3 \ldots \bar{x}_n$) توصیف می شوند و توسط توزیع احتمال ثابت و معلوم $\mathcal{D}$ انتخاب می شوند. می دانیم که مفهوم هدف عطفی از متغیر های تصادفی و عکسشان است (مثلا $x_2 \wedge \bar{x}_5$) و الگوریتم یادگیری از این کلاس مفهوم به عنوان فضای فرضیه ای H استفاده می کند. به یک یادگیر سازگار مجموعه ای از ۱۰۰ نمونه انتخابی با $\mathcal{D}$ داده می شود. این یادگیر فرضیه ی h را از H که با تمامی ۱۰۰ نمونه سازگار است خروجی می دهد. (بدین معنا که خطای h بر روی این ۱۰۰ نمونه صفر است)

(a) علاقه داریم که خطای واقعی h را که احتمال دسته بندی اشتباه نمونه های انتخابی با $\mathcal{D}$ است را بیابیم. بر اساس اطلاعات بالا آیا می توانید بازه ای را مشخص کنید که خطای واقعی حداقل با احتمال ۹۵٪ در آن قرار گیرد؟ اگر چنین است، بازه را بیان کرده و به وضوح آنرا توجیه کنید. اگر امکان پذیر نیست مشکل را توضیح دهید.

(b) حال مجموعه نمونه ی جدید ۱۰۰ تایی دیگری به طور مستقل با همان توزیع $\mathcal{D}$ انتخاب می کنید و معلوم می شود که h ۳۰ نمونه از این ۱۰۰ نمونه را اشتباه دسته بندی می کند. آیا می توانید بازه ای ارائه کنید که این خطای واقعی با احتمال ۹۵٪ در آن قرار داشته باشد؟ (برای این قسمت از کارایی فرضیه روی نمونه های آموزش اش صرف نظر کنید.) اگر چنین است، بازه را بیان کرده و به وضوح آنرا توجیه کنید. اگر امکان پذیر نیست مشکل را توضیح دهید.

(c) ممکن است کمی عجیب به نظر برسد که با این که فرضیه نمونه های آموزشی را درست دسته بندی کرده اما در دسته بندی نمونه هـای 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 الگوریتم های پیشبینی	pool of prediction algorithms
پیچیدگی محاسباتی	Computational complexity
پیچیدگی نمونه ای	Sample complexity
ترکیب G	G-composition
توانی	exponential
چارچوب کران خطا	mistake bound framework
چهارچوب	framework
چهارچوب تقریبا درست	probably approximately correct (PAC)
خرد کردن مجموعه ای از نمونه ها	shattering a set of instances
خطای آموزشی	training error
خطای واقعی	True Error
دقیق	Exact
فرم نرمال فصلی k جمله ای	k-term disjunctive normal form (k-term DNF)
قابلیت یادگیری PAC	PAC-learnability
گراف جهت دار بدون دور	directed acyclic
گراف لایه ای	layered graph
مدل یادگیری تقریبا درست	probably approximately correct (PAC)
مرز بدترین حالت ممکن	wose-case bound
مرز خطا	Mistake bound
مرز های Hoeffding	Hoeffding bounds
مرز های اضافی Hoeffding	Hoeffding additive bounds