نظریه یادگیری ماشین

حمید بیگی

دانشگاه صنعتی شریف

بهار ۱۴۰۱

حق انتشار

درس‌نامه پیش‌رو از روی یادداشت‌های درس نظریه‌یادگیری‌ماشین که در دانشگاه صنعتی شریف تدریس می‌شود نوشته شده است. با توجه به اینکه ممکن است ایرادات نگارشی در متن این درس‌نامه وجود داشته باشد خواهشمند است در صورت وجود هر گونه اشتباهی اعم از نگارشی و یا منطقی خواهشمند است از طریق رایانامه beigy@sharif.edu به آگاهی بنده برسانید. همچنین از انتشار این درس‌نامه به دلیل امکان وجود ایرادهای نگارشی به افراد دیگر بدون آگاهی بنده خوداری نمایید.

فهرست مطالب

فهرست نمادها

۱

دیباچه

۱.۱ یادگیری ماشین چیست؟

یادگیری ماشین مطالعه روش های خودکاری است که پیش بینی های دقیق یا کنش های مناسب را براساس تجربیات گذشته یا مشاهده ها انجام می دهد. الگوریتم های یادگیری ماشین باید در زمان مناسب و با حافظه مناسب قابل اجرا باشند و به حجم داده زیادی نیاز نداشته باشند. به این الگوریتم ها، الگوریتم های کارا[1] می گویند. در چه زمانی به یادگیری ماشین نیاز مندیم؟

۱. وظایف بسیار پیچیده هستند تا بتوان آنها را برنامه نویسی (پیاده سازی) نمود.

(آ) وظایفی که انجام آنها برای انسان ها یا حیوان ها بسیار ساده هستند اما برای رایانه ها بسیار سخت هستند.

(ب) وظایفی که انجام آنها بیشتر ز توان انسان ها است. برای نمونه تحلیل داده های ستاره شناسی، پیش بینی هوا، تحلیل داده های زیستی، تجارت الکترونیک و جویشگرهای وب نمونه ای از وظایف هستند.

۲. بسیاری از وظایف با زمان تغییر می کنند و برنامه های رایانه ای می بایست قابلیت تطبیق داشته باشند تا توانایی کارکردن برای این وظایف را داشته باشند. برای نمونه تشخیص هرزنامه نمونه ای از وظایفی است که نیاز به تطبیق دارد.

یادگیری ماشین دارای چه کاربردهایی هستند؟ برخی از کاربردهای یادگیری ماشین عبارتند از

۱. دسته بندی متون مانند دسته بندی نامه ها

۲. پردازش زبان طبیعی مانند بازشناسی موجودیت های نامدار

[1] Efficient

۳. بازشناسی گفتار و تبدیل گفتار به متن
۴. بازشناسی نوری حروف
۵. کاربردهای زیست شناسی
۶. بینایی ماشین
۷. تشخیص تقلب
۸. بازی های رایانه ای
۹. تشخیص های پزشکی
۱۰. سامانه های تحلیل نظر و پیشنهاد دهنده ها

دسته بندی های مختلف روش های یادگیری ماشین

۱. روش های با نظارت در مقابل روش های بی نظارت، نیمه نظارتی، و یادگیری تقویتی
۲. روش های یادگیری فعال در مقابل روش های یادگیری غیر فعال
۳. روش های یادگیری برخط در مقابل روش های یادگیری دسته‌ای

مسایل یادگیری ماشین را می توان به دسته های مختلفی تقسیم نمود. مهم ترین این دسته بندی ها عبارتند از

۱. دسته بندی
۲. رگرسیون
۳. رتبه بندی
۴. خوشه بندی
۵. کاهش بعد
۶. یادگیری متغیرهای پنهان

۲.۱ اهداف پژوهش های یادگیری ماشین

۱. کارا باشند
۲. پاسخ های تولید شده تفسیر پذیر باشند
۳. پاسخ تولید شده تا حد ممکن دقیق باشد

۳.۱ نظریه یادگیری ماشین

هدف نظریه یادگیری ماشین توسعه و تحلیل مدل های رسمی است که به فهم مسایل مهم در یادگیری ماشین کمک کند. این نظریه کمک می کند تا بفهمیم کدام مفاهیم را می توان به صورت کارا یادگرفت و چه اندازه داده برای یادگیری این مفاهیم مورد نیاز است. همچنین این نظریه تلاش می کند تا پاسخ پرسشی که چرا یک الگوریتم در همه شرایط خوب کار نمی کند را بیابد. برخی از پرسش های مهم که در نظریه یادگیری ماشین مطرح هستند عبارتند از

۱. تعداد نمونه مورد نیاز برای یک یادگیری موفق
۲. پیچیدگی محاسباتی یک مساله و/یا الگوریتم یادگیری
۳. پیدا نمودن پیچیدگی محاسباتی یک الگوریتم یادگیری براساس تعداد نمونه های آموزشی
۴. دسته بندی مسایل یادگیری به مسایل ساده و سخت.
۵. چگونگی به کار بردن اطلاعات زمینه و پیشین در مورد مساله یادگیری برای یادگیری موثر
۶. یافتن پاسخ برای بهتر بودن فرضیه های ساده تر
۷. چگونگی تغییر کارایی یادگیر بر اساس چگونگی تحویل داده های آموزشی
۸. اگر یادگیر نمونه آموزشی درخواست نماید پیچیدگی نمونه ای یادگیر چگونه تغییر می یابد؟
۹. آیا داده های برچسب نخورده می توانند کمکی به آموزش / افزایش کارایی الگوریتم یادگیری نمایند؟
۱۰. اگر یادگیری به صورت توزیع شده صورت پذیرد هزینه ارتباطی چه اندازه خواهد بود؟
۱۱. چگونه می توان از یادگیر ضعیف استفاده نمود؟

۲

مدل های رسمی یادگیری

برای تحلیل و مطالعه الگوریتم های یادگیری ماشین نیاز است تا مساله یادگیری به صورت رسمی و دقیق تعریف شود. این تعریف دقیق، مدل یادگیری نام دارد. مدل یادگیری باید به اندازه کافی دقیق باشد و توانایی مدل سازی وجه‌های مهم مسایل یادگیری واقعی را داشته باشد. دقت مدل به این دلیل مورد نیاز است که بتوان مساله یادگیری را به صورت ریاضی مطالعه نمود. همچنین یک مدل یادگیری خوب میبایست در مقابل تغییرات جزیی در تعریف آن مقاوم باشد. برای هر مدل رسمی فرض‌های ساده کننده غیرقابل اجتناب است. مدل‌های یادگیری می‌بایست توانایی پاسخ به پرسش‌های گوناگون همانند پرسش‌های زیر را داشته‌باشد.

۱. چه چیزی باید یادگرفته شود؟
۲. داده ها چگونه تولید میشوند؟
۳. داده ها چگونه به یادگیر داده میشوند؟ برای نمونه داده ها یکی یکی یا به صورت دسته ای به یادگیر داده میشوند؟
۴. هدف از یادگیری در این مدل چیست؟

در ادامه این بخش، نخست به بررسی مولفه‌های مدل‌های رسمی یادگیری می‌پردازیم و سپس مدل‌های مختلف یادگیری بررسی و نقاط ضعف و قوت آنها تحلیل می‌گردند.

۱.۲ مولفه های مدل رسمی یادگیری

مدل‌های رسمی یادگیری دارای پنج مولفه اصلی هستند: ورودی یادگیر، خروجی یادگیر، مدل تولید داده، معیار کارایی و دانش پیشین. در ادامه هر کدام از این مولفه ها شرح داده می‌شوند.

۱. یکی از مولفه‌های هر مدل یادگیری، ورودی یادگیر است. در مدل رسمی یادگیری، یادگیر به داده های زیر دسترسی دارد.

- فضای نمونه که با $\mathcal{X}$ نشان داده می‌شود عبارت است از مجموعه نمونه‌هایی که ممکن است در آینده برچسب آنها تعیین شوند. به هریک از اعضا درون $\mathcal{X}$ یک نمونه می‌گویند. یک نمونه به وسیله‌ی مجموعه‌ای از خصوصیات یا ویژگی ها توصیف شده و معمولا با یک بردار n بعدی بازنمایی می‌شود.
- مجموعه برچسب ها که با $\mathcal{Y}$ نشان داده می‌شود عبارت است از برچسب یا گروهی که می‌خواهیم یادگیر آن را پیش بینی نماید. برای نمونه در تشخیص اعداد، برچسب نمونه ها از مجموعه $\mathcal{Y} = \{۰, ۱, ۲, ۳, ۴, ۵, ۶, ۷, ۸, ۹\}$ انتخاب می‌شوند و همچنین $\mathcal{Y} = \{۰, ۱\}$ و $\mathcal{Y} = \{-۱, +۱\}$ دو نمونه از مجموعه برچسب های ممکن برای دسته‌بندی دودویی هستند.
- در هنگام آموزش یادگیر، مجموعه ای از نمونه های برچسب دار که مجموعه آموزشی نامیده میشود به عنوان ورودی به الگوریتم یادگیری داده می‌شود و در هنگام آزمایش، تنها نمونه های آزمایشی که برچسب ندارند به یادگیر داده می‌شود و یادگیر می‌بایست برچسب آنها را تعیین نماید. مجموعه آموزشی به صورت $\mathsf{S} = \{(x_۱, y_۱), (x_۲, y_۲), ..., (x_m, y_m)\}$ نشان داده می‌شود که دنباله‌ای از زوج‌های مرتب $(x, y) \in \mathcal{X} \times \mathcal{Y}$ است. به این مجموعه نقاط برچسب خورده دامنه نیز گفته می‌شود.

۲. مولفه دیگر مدل رسمی یادگیر، خروجی آن است. یک یادگیر می‌بایست یک خروجی به شکل $h : \mathcal{X} \mapsto \mathcal{Y}$ تولید نماید. به این خروجی فرضیه یا در مسایل دسته‌بندی، دسته بند می‌گویند. این فرضیه بعدا برای پیش بینی برچسب نمونه های آزمایشی (نمونه‌های برچسب نخورده) که از فضای $\mathcal{X}$ با توزیع $\mathcal{D}$ نمونه برداری شده‌اند به کار می‌رود. در اینجا نماد $A(\mathsf{S})$ را برای نشان دادن فرضیه ای که الگوریتم یادگیری A با دریافت مجموعه آموزشی S تولید می‌نماید به کار می‌بریم. در واقع یادگیری به نوعی ایجاد یک الگوریتم A است که اگر به آن مجموعه آموزشی S را بدهیم به ما h را به عنوان خروجی تحویل خواهد داد.

۳. شیوه نمونه برداری نمونه ها یا به عبارتی دیگر مدل تولید داده ها بخش دیگری از مدل رسمی یادگیر است. در این‌جا فرض می‌کنیم داده‌های آموزشی و آزمایشی از یک توزیع احتمال $\mathcal{D}$ از فضای نمونه‌ها، نمونه برداری شده باشند و یادگیر هیچ گونه اطلاعی از این توزیع ندارد. برای ایجاد مجموعه آموزشی، پس از نمونه برداری هر نمونه x، برچسب آن توسط تابع $c : \mathcal{X} \mapsto \mathcal{Y}$ تعیین میشود. به تابع c مفهوم می‌گویند و آموزگار آن را می‌داند. یادگیر شیوه برچسب زنی آموزگار یا همان تابع c را نمی‌داند. اغلب فرض می‌کنیم که نمونه ها با یک مفهوم ناشناخته که متعلق به یک کلاس مفاهیم شناخته شده‌است برچسب زده می‌شود.

۴. مولفه دیگر مدل رسمی یادگیری، معیار ارزیابی کارایی فرضیه تولید شده‌است. کارایی یک فرضیه را می‌توان بر اساس معیارهای متفاوتی ارزیابی نمود. یکی از این معیارها خطای واقعی‌فرضیه است که به صورت زیر تعریف میشود. خطای واقعی فرضیه h برابر است با احتمال اینکه نمونه تصادفی x که با توزیع $\mathcal{D}$ نمونه برداری شده است به فرضیه h داده شود و برچسب فرضیه h با برچسب واقعی آن یکسان نباشد. به عبارتی دیگر خطای فرضیه h به صورت زیر تعریف میشود.

$$\mathbf{R}(h) = \mathop{\mathbb{P}}_{x \sim \mathcal{D}} [h(x) \neq c(x)] \tag{۱.۲}$$

این خطا با نام‌های دیگری مانند خطای تعمیم ، ریسک و خطای واقعی نام برده می‌شود و معمولا به جای همدیگر به کار می‌روند.

۵. علاوه بر مولفه های یاد شده، یک یادگیر به دانش پیشین در فرایند یادگیری نیازمند است. یادگیر اطلاعاتی در باره توزیع $\mathcal{D}$ یا تابع c ندارد و تنها به مجموعه آموزشی دسترسی دارد. ممکن است برای یادگیری فرضیاتی در باره مساله داشته باشیم که به آنها Inductive Bias می‌گویند. بر اساس این فرضیات تعیین شده، الگوریتم یادگیری طراحی می‌شود.

در ادامه این بخش برخی از بخش ها که پیش از این به آنها اشاره شد به صورت رسمی تعریف می‌گردد.

تعریف ۱.۲ مفهوم

یک مفهوم یک تابع دودویی در فضای نمونه است و به صورت زیر تعریف می‌گردد.

$$c : \mathcal{X} \mapsto \{۰, ۱\} \tag{۲.۲}$$

تابع c یا همان مفهوم ناشناخته است و آموزگار این تابع را می‌داند و بر اساس این تابع نمونه های دریافت شده از فضای نمونه رابرچسب می‌زند. برای نمونه مفهوم $c(x_۱, x_۲, \ldots, x_n) = x_۱ \wedge x_۲$ یک تابع دودویی است و مقدار آن برای یک نمونه زمانی یک است که مقدار دو ویژگی نخست نمونه برابر یک باشد. در بیشتر مواقع از واژه های *دسته*، *بند*، *فرضیه* و *قاعده پیش بینی* برای معنای یک مفهوم به کار می‌بریم. برای نمونه در فضای نمونه حیوانات، تفکیک حیوانات از هم یک کلاس مفاهیم را تشکیل می‌دهد و تشخیص سگ از بقیه حیوانات یک مفهوم است.

تعریف ۲.۲ کلاس مفاهیم

منظور از کلاس مفاهیم مجموعه‌ای از مفاهیم همراه با بازنمایی متناظر آن است.

بازنمایی یک کلاس از مفاهیم به این علت مهم است که اندازه‌ی مفاهیم به کمک آن به دست می‌آید.

تعریف ۳.۲ اندازه مفهوم

منظور از اندازه مفهوم عبارت است از تعداد بیت‌های لازم برای بازنمایی مفهوم در بازنمایی تعریف شده آن.

برای روشن شدن موضوع، در ادامه یک مثال از یک مفهوم به همراه بازنمایی‌های مختلف آن و اندازه مفهوم در هر کدام از این بازنمایی‌ها بیان می‌شود.

مثال ۱.۲ (ترکیب عطفی یکنوا).
کلاس عطف یکنوا را درنظر بگیرید. برای نمونه $c(x_۱, x_۲, ..., x_n) = x_i \wedge x_j$ یکی از اعضای این کلاس است. در این توابع نقیض متغیرها را نداریم. این کلاس را می‌توان به راه‌های متفاوتی بازنمایی داد که در ادامه به چند نمونه از آن‌ها می‌پردازیم.

۱. یک راه بازنمایی این کلاس مفاهیم استفاده از یک آرایه‌ی n بیتی است به نوعی که اگر x_i در مفهوم وجود داشته باشد، بیت iام برابر مقدار ۱ و در غیر اینصورت برابر ۰ است. در این حالت اندازه‌ی مفهوم از مرتبه $O(n)$ خواهد بود.
۲. یک راه دیگر بازنمایی این کلاس از مفاهیم این است که به ازای هر ورودی ممکن برچسب آن (۰ یا ۱ بودن) را درنظر بگیریم. در این حالت با توجه به اینکه ۲^n ورودی ممکن خواهیم داشت اندازه‌ی مفهوم نیز از مرتبه $O(۲^n)$ خواهد بود.
۳. راه دیگر این است که زیرنویس و اندیس متغیرهای استفاده شده در ترکیب عطفی را به شکل عددهای $\lg n$ بیتی در کنار هم بنویسیم. در این حالت اگر مفهوم مورد نظر k متغیر عطفی داشته باشد، آنگاه اندازه‌ی مفهوم از مرتبه $O(k \times \lg n)$ خواهد بود.

لذا روشن است که انتخاب شیوه مناسب بازنمایی مفاهیم، اندازه مفهوم را تغییر می‌دهد و در هنگامی که الگوریتم یادگیری تابعی از اندازه مفهوم باشد شیوه بازنمایی مفهوم در کارایی الگوریتم یادگیری تاثیر دارد.

تمرین ۱.۲ (اندازه مفهوم درخت تصمیم) اندازه مفهوم کلاس مفاهیم درخت تصمیم در فضای $\{۰, ۱\}^n$ چه اندازه است؟

۲.۲ مدل سازگاری

در ادامه این بخش، نخستین مدل رسمی یادگیری بیان میشود. این مدل بر اساس سازگاری فرضیه تولید شده توسط الگوریتم یادگیری بنا شده است. این مدل بسیار ساده است و توانایی بکارگیری در کاربردهای واقعی را نداشته اما برای آغاز بررسی مدل‌های یادگیری، مدل بسیار مناسبی است.

تعریف ۴.۲ فرضیه‌ی سازگار

فرضیه‌ی h با داده‌های آموزشی سازگار است اگر و تنها اگر برای همه‌ی نمونه‌های مجموعه آموزشی، داشته باشیم

$$h(x) = c(x) \qquad \forall x \in \mathcal{S} \tag{۳.۲}$$

یا به عبارتی دیگر برای همه نمونه های آموزشی، برچسب تولید شده توسط فرضیه h با برچسب اصلی آن یکسان باشد.

بر اساس تعریف فرضیه های سازگار، مدل سازگاری به صورت زیر تعریف میشود.

تعریف ۵.۲ یادگیری در مدل سازگاری

الگوریتم یادگیری A کلاس مفاهیم C را در مدل سازگاری یاد می‌گیرد، اگر برای هر مجموعه‌ی آموزشی $\mathcal{S}$، الگوریتم A مفهوم $c \in C$ که سازگار با $\mathcal{S}$ باشد را در صورت وجود تولید و در غیر اینصورت پاسخ "فرضیه‌ای وجود ندارد" را تولید می‌نماید.

تعریف ۶.۲ یادگیری کارا در مدل سازگاری

کلاس مفاهیم C در مدل سازگاری به صورت *کارا* یادگرفتنی است اگر یک الگوریتم *کارای* A وجود داشته باشد که کلاس مفاهیم C را در مدل سازگاری یاد بگیرد. کارایی براساس اندازه‌ی مجموعه $\mathcal{S}$ که با m نشان داده می‌شود و اندازه‌ی نمونه‌های ورودی که با n نشان داده‌می‌شود تعریف می‌شود.

الگوریتم کارا معمولاً به الگوریتمی گفته می‌شود که در زمان چندجمله‌ای براساس اندازه S و اندازه x مساله را حل می‌کند.

مثال ۲.۲ (توابع عطفی یکنوا).
این کلاس مفاهیم شامل تمامی مفاهیمی است که می‌توان از ترکیب عطفی یکنوا حداکثر n متغیر دودویی ساخت. در این ترکیب عطفی، نقیض متغیرها ظاهر نمی‌شوند. برای نمونه، مجموعه آموزشی زیر که در آن یک تابع عطفی سازگار وجود دارد را در نظر بگیرید.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	y
1	0	1	1	0	0	1	1	+
1	1	1	1	1	0	1	0	+
0	1	1	1	0	0	1	1	+
0	0	0	1	1	1	1	1	−
1	1	1	1	1	0	0	0	−

الگوریتم زیر را برای حل این مثال درنظر بگیرید :

۱. در نمونه‌های مثبت، تمام ویژگی‌هایی که دارای مقدار صفر هستند را حذف کن.

۲. ترکیب عطفی ویژگی‌های باقیمانده را حساب کن.
۳. اگر فرضیه تولید شده با نمونه‌های منفی هم سازگار بود این فرضیه را به عنوان خروجی اعلام و در غیر اینصورت پاسخ "فرضیه‌ای وجود ندارد" را تولید کن.
با مجموعه آموزشی یاد شده در بالا، حاصل اعمال الگوریتم فوق $c(x) = x_3 \wedge x_4 \wedge x_7$ است. با توجه به این‌که این الگوریتم در زمان $O(nm)$ خروجی می‌دهد، بنابراین این الگوریتم کارا است.

درنتیجه این کلاس مفاهیم به‌صورت کارا یادگرفتنی است و در نتیجه قضیه زیر را خواهیم داشت.

قضیه ۱.۲ (یادگیری ترکیب عطفی یکنوا در مدل سازگاری).
کلاس مفاهیم ترکیب عطفی یکنوا در مدل سازگاری به صورت کارا یادگرفتنی است.

تمرین ۲.۲ (یادگیری ترکیب عطفی غیر یکنوا) کلاس مفاهیم ترکیب عطفی غیر یکنوا عبارت است از ترکیب عطفی تعدادی متغیر یا نقیض آنها. نشان دهید که این کلاس مفاهیم بصورت کارا یادگرفتنی هستند و برای یادگیری این کلاس مفاهیم یک الگوریتم یادگیری طراحی نموده و کارایی آن را بررسی نمایید.

برای حل این مساله نقیض هر متغیر را با یک متغیر جدید نشان می‌دهیم و مساله را به یادگیری ترکیب عطفی یکنوا تبدیل میکنیم. برای نمونه تابع اگر هدف یادگیری یک ترکیب عطفی غیریکنوا روی n متغیر است. با تعریف $z_i = \not{x}_i$، مساله به یادگیری ترکیب عطفی با $2n$ متغیر تبدیل می‌شود و از الگوریتمی که پیش از این ارایه شد می‌توان استفاده نمود.

مثال ۳.۲ (ترکیب فصلی یکنوا).
این کلاس مفاهیم از ترکیب فصلی حداکثر n متغیر (ویژگی) دودویی تولید میشود. در این ترکیب فصلی، نقیض متغیرها ظاهر نمی‌شوند. برای نمونه تابع $c(x_1, x_2, \ldots, x_n) = x_2 \vee x_6 \vee x_8$ یک ترکیب فصلی است. به سادگی میتوان نشان داد که با کمک قانون دمورگان، هر ترکیب فصلی را میتوان به ترکیب عطفی تبدیل نمود. برای نمونه اگر $c(x) = x_i \vee x_j$ باشد، می‌توان آن را به صورت ترکیب عطفی زیر نیز درآورد $c(x) = x_i \vee x_j = \overline{\overline{x_i} \wedge \overline{x_j}}$ که $\overline{x_i}$ یعنی نقیض x_i است. یعنی اگر مقدار x_i برابر ۰ باشد $\overline{x_i}$ برابر ۱ خواهد بود و اگر مقدار x_i برابر ۱ باشد $\overline{x_i}$ برابر ۰ خواهد بود. همچنین علامت تغییر مقدار که بالای مقدار $\overline{x_i} \wedge \overline{x_j}$ قرار گرفته است به معنی تغییر برچسب است.
فرض کنید می‌خواهیم کلاس ترکیب فصلی را یاد بگیریم. الگوریتمی که ارائه می‌دهیم با کمی تغییر الگوریتم قبلی به دست می‌آید. بدین ترتیب که ابتدا برچسب‌ها و بیت‌ها را وارون می‌کنیم (یعنی اگر مقدار ۰ داشتند مقدارشان را ۱ می‌کنیم و اگر مقدار ۱ داشتند مقدارشان را ۰ درنظر می‌گیریم) و سپس با استفاده از الگوریتم قبلی جواب را به دست می‌آوریم و در نهایت همه چیز را به حالت اولیه برمی‌گردانیم.

مثال ۴.۲ (kمساله‌ی فرم نرمال عطفی درجه).
مساله‌ی kفرم نرمال عطفی درجه را درنظر بگیرید. اگر ترکیب فصلی حداکثر k متغیر را یک «عبارت» بنامیم، آنگاه مسئله‌ی فرم نرمال عطفی درجه k[a] عبارت است از ترکیب عطفی تعدادی «عبارت». یادگیری در این مساله بدین معنی است که k داده شود وپرسش این باشد که آیا میتوان کلاس مفاهیمی که با فرم نرمال عطفی درجه k نشان داده میشود را یادگرفت یا خیر؟ ما فرض میکنیم که k یک عدد کوچک باشد. ما میتوانیم این مساله را به یادگیری ترکیب عطفی کاهش دهیم. برای این کار متغیر جدیدی برای هر عبارت ممکن در فرم نرمال عطفی درجه k را ایجاد میکنیم.
در حالت کلی متغیرهای z_{s_i} را از روی ترکیب فصلی s تا از متغیرهای اولیه‌ی x_j ها می‌سازیم. بدین ترتیب تمامی حالات s متغیری را در z_{s_i} ایجاد می‌نماییم. مثلاً $z_{s_i} = \{x_{j_1} \vee x_{j_2} \vee \ldots \vee x_{j_s}\}$ برابر i امین متغیر جدید z است که دقیقاً از ترکیب فصلی s متغیر ایجاد شده است. برای نمونه اگر $k = 2$ باشد برای کلاس مفاهیم فرم نرمال عطفی درجه k متغیرهای زیر را

خواهیم داشت.

$$\begin{aligned} z_۱ &= (x_۱) \\ z_۲ &= (x_۲) \\ z_۳ &= (x_۱ \vee x_۲) \\ z_۴ &= (x_۱ \vee \overline{x_۲}) \\ z_۵ &= (\overline{x_۱} \vee x_۲) \\ z_۶ &= (\overline{x_۱} \vee \overline{x_۲}) \end{aligned}$$

اگر تعداد متغیرهای جدید ایجاد شده را P بنامیم، خواهیم داشت

$$P = \sum_{i=۱}^{k} \binom{۲n}{i} \leq (۲n)^k \tag{۴.۲}$$

بنابراین اگر تعداد نمونه‌ها را m عدد فرض کنیم، الگوریتم ذکر شده از $O(m(۲n)^k)$ است که برای k های کوچک، مثلاً ۲ یا ۳، الگوریتم کارا است و در غیراینصورت الگوریتم کارا نخواهد بود.

[a] k-CNF

در این مثال، نمونه های ۲ بیتی به نمونه های ۶ بیتی تبدیل میشوند که در آن بیت iام در صورتی برابر یک است اگر z_i برابر یک باشد و در غیر اینصورت برابر با صفر خواهد بود. اگر نمونه های آموزشی را به الگوریتم یادگیری که برای یادگیری ترکیب عطفی بدهیم مفهوم سازگار با داده های آموزشی را بر اساس z_i تولید خواهد نمود. سپس با جایگزینی z_i به ترکیب فصلی تعریف شده آن، مفهوم مساله‌ی فرم نرمال عطفی درجه k سازگار را پیدا میکند. برای یادگیری این کلاس مفاهیم در حالت کلی نیاز به تعریف $O((۲n)^k)$ متغیر z_i نیاز است. این مقدار براساس رابطه زیر بدست میآید که تعداد فرم نرمال عطفی درجه kهایی که از n متغیر می‌توان تولید نمود برابر است با

$$(۲n)(۲n-۱)(۲n-۲)\dots(۲n-k) = O\left((۲n)^k\right). \tag{۵.۲}$$

زیرا هر مکان در فرم نرمال عطفی درجه k را می‌توان از $۲n$ متغیر (خود یا نقیض آن) استفاده نمود. بنابراین اگر k عددی کوچک باشد ، این الگوریتم کارا است و اگر k بزرگ باشد زمان اجرای الگوریتم نمایی خواهد بود.

مثال ۵.۲ (کلاس مفاهیم DNF (Or of ands)).

فرض کنید می‌خواهیم کلاس مفاهیم DNF را یاد بگیریم. این کلاس به صورت ترکیب فصلی ($\vee$) تعدادی عبارت که هرکدام به صورت ترکیب عطفی ($\wedge$) تعدادی متغیر است است. برای نمونه

$$c(x_۱, x_۲, \dots, x_n) = (x_۲ \wedge x_۷ \wedge x_{۱۰}) \vee (x_۱ \wedge x_۵) \vee (x_۳ \wedge x_۴ \wedge x_۶) \tag{۶.۲}$$

یک عبارت DNF است. برای یادگیری این کلاس مفاهیم میتوانیم از الگوریتم زیر استفاده کنیم.

۱. ترکیب فصلی نمونه های مثبت را محاسبه کن.
۲. اگر این ترکیب فصلی با نمونه های منفی سازگار بود آن را بعنوان خروجی و در غیر اینصورت پاسخ "فرضیه‌ای وجود ندارد" را تولید کن.

این الگوریتم کارا است و در زمان $O(nmk)$ قابل پیاده‌سازی است که k حداکثر تعداد عبارت‌های هر نمونه است. بنابراین این کلاس نیز در مدل سازگاری بصورت کارا یادگرفتنی است. اما فرضیه تولید شده کاربرد ندارد زیرا

۱. این فرضیه داده های مثبت را ذخیره نموده و بیش برازش رخ میدهد و هر نمونه ای که در مجموعه آموزشی وجود نداشته باشد را به

عنوان نمونه منفی برچسب میزند. این ایراد اختلاف بین قدرت حافظه و قدرت تعمیم فرضیه را نشان می‌دهد.

۲. اندازه فرضیه تولید شده برابر است با مجموع اندازه نمونه های مثبت مجموعه آموزشی.

مدل سازگاری، مدلی بسیار ساده است اما در عمل کاربرد ندارد زیرا این مدل دارای ایرادهایی است. برخی از ایرادهای کلی و عمومی مدل سازگاری در زیر آمده است.

۱. در این مدل هیچ سخنی در باره قدرت تعمیم فرضیه تولید شده به میان نیامده است. یک مدل یادگیری خوب می‌بایست در مورد توان (قابلیت) فرضیه تولید شده روی داده های جدید سخن بگوید.

۲. مشکل دوم این مدل این است که اندازه‌ی مفاهیم به میزان اندازه‌ی ورودی‌ها (و طبعاً بزرگ‌تر از حالت ایده‌آل) هستند

۳. ایراد دیگر این مدل این است که این مدل به نویز مقاوم نیست. با توجه به اینکه در کاربردهای واقعی داده ها نویزی هستند و به علت وجود داده‌های نویزی، خطای آموزشی صفر نمی‌شود لذا این مدل در عمل قابل استفاده نیست.

۴. و آخرین مشکلی که در این‌جا بیان می‌کنیم این است که اگر برای کلاس مفاهیم C الگوریتم کارا وجود داشته باشد، برای زیرمجموعه‌های C مانند C' که $C' \subset C$ لزوماً نمی‌توان الگوریتم کارا پیدا نمود. برای بیان بهتر این ایراد به مثال زیر توجه نمایید.

مثال ۶.۲ (یادگیری کلاس مفاهیم فرم نرمال فصلی دو لفظی).
فرض کنید که کلاس مفاهیم C شامل همه فرم نرمال فصلی دو لفظی (یا دو ترکیب عطفی با طول دلخواه) باشد. برای نمونه تابع $c(x_1,\ldots,x_n) = (x_2 \wedge x_7 \wedge x_{10}) \vee (x_1 \wedge x_5)$ یک فرم نرمال فصلی دو لفظی [a] است. حال چگونه میتوانیم یک فرم نرمال فصلی دو لفظی از روی داده های آموزشی یاد بگیریم؟ اگر بتوانیم یادگیری مساله فرم نرمال فصلی دو لفظی را به مساله یادگیری فرم نرمال عطفی درجه k تبدیل کنیم کار ساده تر میشود. اگر از قواعد منطق (با در نظر گرفتن $\wedge$ معادل جمع و $\vee$ معادل ضرب) استفاده کنیم می‌توانیم این تبدیل را انجام بدهیم. برای نمونه

$$c(x_1,\ldots,x_n) = (x_1 \wedge x_2) \vee (x_3 \wedge x_4) = (x_1 \vee x_3) \wedge (x_1 \vee x_4) \wedge (x_2 \vee x_3) \wedge (x_2 \vee x_4).$$

این مثال نشان می‌دهد که ما همیشه می‌توانیم فرم نرمال فصلی دو لفظی را به فرم نرمال عطفی درجه k تبدیل نماییم. آیا این به معنای این است که این کلاس مفاهیم یادگرفتنی است؟
اگر بتوانیم برعکس تبدیل بالا را یادبگیریم مساله ساده می‌شود. با تبدیل فرم نرمال فصلی دو لفظی به فرم نرمال عطفی درجه k می‌توانیم فرم نرمال عطفی درجه k را یاد بگیریم و سپس باید پاسخ فرم نرمال عطفی درجه k تولید شده را بتوانیم به فرم نرمال فصلی دو لفظی تبدیل نماییم. مشکل اصلی این است که ما همیشه می‌توانیم یک فرم نرمال فصلی دو لفظی را به فرم نرمال عطفی درجه k تبدیل نماییم اما الزاما نمی‌توانیم یک فرم نرمال عطفی درجه k را به یک فرم نرمال فصلی دو لفظی تبدیل نماییم. در نتیجه اگر ما یک فرم نرمال عطفی درجه k سازگار را پیدا کنیم بدان معنی نیست که یک فرم نرمال فصلی دو لفظی سازگار وجود داشته باشد. یا بطور غیر رسمی فضای فرم نرمال عطفی درجه k یک ابر فضا از فضای فرم نرمال فصلی دو لفظی است و یا به عبارتی دیگر

$$(۲ - termDNF) \subseteq (۲ - CNF). \qquad (۷.۲)$$

در زمینه نظریه یادگیری، یادگیری فرم نرمال عطفی درجه k آسان است اما یادگیری کلاس مفاهیم به شکل فرم نرمال فصلی دو لفظی یک مساله انپی‌سخت است.

[a] 2-term DNF

۳.۲ مدل یادگیری احتمالا تقریبا درست

در بخش پیش مدل سازگاری بیان گردید. همچنین درباره دو کاستی بزرگ این مدل سخن به میان آمد. نخست آن که این مدل تعریفی از

قدرت تعمیم فرضیه ها ارایه نمی‌کند و دوم آنکه این مدل امکان پردازش داده‌های نویزی را ندارد. این دو کاستی سبب غیر کاربردی شدن این مدل می‌شود. در این بخش تلاش می‌کنیم که مدل را توسعه دهیم تا در تعریف مدل قدرت تعمیم فرضیه ها ارایه شود و پردازش داده های نویزی را در آینده بررسی خواهیم‌نمود. در این بخش مدل احتمالا تقریبا درست بیان می‌شود که در آن قدرت تعمیم فرضیه ها در مدل گنجانده شده‌است. برای بررسی قدرت تعمیم فرضیه ها فرض‌های ساده کننده زیر را در نظر می‌گیریم که بیشتر این فرض‌ها در آینده تعدیل خواهندشد.

۱. فرض می‌کنیم که نمونه‌های آزمون و آموزش از توزیع ثابت و یکسان اما ناشناخته $\mathcal{D}$ به صورت مستقل نمونه برداری می‌شوند. این توزیع برای یادگیر ناشناخته است.

۲. نمونه ها با کمک یک تابع ثابت و ناشناخته $c \in \mathcal{C}$ برچسب زده می‌شوند. یادگیر نمی‌داند که این تابع چیست و دارای چه شکلی است.

در ادامه نخست خطای واقعی فرضیه h که نشان دهنده قدرت تعمیم آن است را تعریف می‌کنیم.

تعریف ۷.۲ **خطای واقعی**

خطای واقعی هر فرضیه $h \in H$ که خطای تعمیم آن نیز نامیده میشود به صورت زیر تعریف می‌شود.

$$\begin{aligned} \mathbf{R}(h) &= \mathop{\mathbb{P}}_{x\sim\mathcal{D}}[c(x) \neq h(x)] \\ &= \sum_{c(x)\neq h(x)} \mathcal{D}(x). \end{aligned} \tag{۸.۲}$$

هدف آرمانی الگوریتم‌های یادگیری این است که فرضیه‌ای را پیدا نمایند که خطای واقعی آن را برابر صفر باشد یعنی همه نمونه هایی که در هنگام آموزش ندیده‌است را بدون خطا برچسب بزند. اما این کار به دو دلیل زیر ممکن نیست. (۱) مجموعه آموزشی یک زیرمجموعه از فضای نمونه است و ممکن است چندین فرضیه سازگار با داده‌های آموزشی داشته‌باشیم و هیچکدام از آنها با مفهوم اصلی (c) یکسان نباشند. (۲) احتمال اینکه نمونه های آموزشی از همه بخش‌های فضای نمونه، نمونه برداری نشده‌باشند بزرگتر از صفر است. لذا هیچگاه امکان تولید فرضیه‌ای با خطای واقعی صفر وجود ندارد و به همین جهت خطای واقعی صفر را تعدیل نموده و به جای آن خطای واقعی کوچکتر از ϵ را در نظر می‌گیریم و از آنجایی که داده‌ها از یک فضا به‌وسیله توزیع $\mathcal{D}$ نمونه‌برداری می‌شوند یادگیر می‌بایست برای هر ترتیب از داده‌های آموزشی فرضیه با خطای کم را تولید نماید لذا احتمال شکست δ را در نظر گرفته و تلاش می‌کنیم که احتمال شکست کوچکتر از δ باشند. اگر خطای فرضیه h کوچک باشد فرضیه h را یک فرضیه تقریبا درست [1] می‌گویند. اما همیشه نمی‌توان تضمین نمود که خطای واقعی کوچک باشد زیرا ممکن است داده های آموزشی نماینده خوبی از مفهوم واقعی نباشند به همین جهت باید تعریف بازنگری شده و به دنبال فرضیه‌ای باشیم که با احتمال زیاد تقریبا درست باشد یا به عبارتی دیگر فرضیه h احتمالا تقریبا درست باشد. درستی تقریبی یک فرضیه با کمک پارامتر دقت (ϵ) بیان می‌گردد و به عبارتی دیگر یک فرضیه h تقریبا درست است اگر شرط $\mathbf{R}(h) \leq \epsilon$ برقرار باشد و احتمال درستی با کمک پارامتر اطمینان ($1-\delta$) بیان می‌گردد و به عبارتی فرضیه h را احتمالا تقریبا درست می‌گویند اگر شرط $\mathbb{P}[\mathbf{R}(h) \leq \epsilon] \geq 1-\delta$ برقرار باشد براساس این تعاریف، مدل یادگیری احتمالا تقریبا درست به صورت زیر تعریف می‌شود.

[1] Approximately correct

تعریف ۸.۲ **قابلیت یادگیری احتمالا تقریبا درست**

کلاس مفاهیم $\mathcal{C}$ قابلیت یادگیری احتمالا تقریبا درست را دارد اگر یک الگوریتم A و یک تابع چند جمله ای p وجود داشته باشد به گونه ای که برای هر $\epsilon \in (۰, \frac{۱}{۲})$ و هر $\delta \in (۰, \frac{۱}{۲})$ و برای هر توزیع $\mathcal{D}$ روی X و برای تمامی مفاهیم $c \in \mathcal{C}$ و برای همه مجموعه‌های آموزشی S با انداز $m \geq m_H(\epsilon, \delta) = p\left(\frac{۱}{\epsilon}, \frac{۱}{\delta}, |x|, |C|\right)$ که بصورت مستقل از توزیع $\mathcal{D}$ نمونه برداری شده‌اند رابطه زیر برقرار باشد

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}[\mathbf{R}(h) \leq \epsilon] \geq ۱ - \delta. \tag{۹.۲}$$

اگر الگوریتم یادگیری A در زمان $p\left(\frac{۱}{\epsilon}, \frac{۱}{\delta}, |x|, |C|\right)$ اجرا شود آنگاه میگوییم که کلاس مفاهیم $\mathcal{C}$ به صورت کارا قابل یادگیری است و اگر الگوریتم یادگیری L وجود داشته باشد به آن الگوریتم یادگیری احتمالا تقریبا درست برای کلاس مفاهیم $\mathcal{C}$ میگویند. در این مدل اگر فرض کنیم که یک $h \in H$ وجود دارد به گونه ای که خطای واقعی آن صفر است ($\mathbf{R}(h) = ۰$) یعنی خطای خطای تجربی آن که به صورت زیر تعریف میشود نیز صفر است.

$$\hat{\mathbf{R}}(h) = \frac{۱}{m}\sum_{i=۱}^{m} \mathbb{I}\left[h(x_i) \neq c(x_i)\right]. \tag{۱۰.۲}$$

مدل یادگیری احتمالا تقریبا درست دارای ویژگی‌های بسیار جالبی است که عبارتند از (۱) این چارچوب مستقل از توزیع نمونه برداری $\mathcal{D}$ است. (۲) نمونه‌های آموزشی و آزمایشی که خطا را مشخص می‌کنند از یک توزیع نمونه برداری می‌شوند. (۳) این چارچوب برای یادگیری یک کلاس مفاهیم C است نه یک مفهوم $c \in C$. اگر کلاس مفاهیم C قابلیت یادگیری احتمالا تقریبا درست را داشته باشد آنگاه (۱) تعداد نمونه‌های آموزشی مورد نیاز جهت تولید فرضیه h یک تابع چند جمله‌ای براساس پارامترهایش است و (۲) اگر الگوریتم یادگیری برای هر نمونه زمان ثابتی را مصرف کند آنگاه زمان پردازش برای هر مجموعه آموزشی یک تابع چند جمله‌ای است. در این مدل سه پارامتر ϵ, δ, m وجود دارد و دو پارامتر از این سه پارامتر را مشخص می‌کنیم و در مسایل یادگیری مختلف به دنبال یافتن پارامتر سوم هستیم. اگر ϵ و δ مشخص شده باشند ما به دنبال پیدا کردن اندازه مجموعه آموزشی هستیم که بتواند یک یادگیر احتمالا تقریبا درست با پارامترهای مشخص شده را تولید نماید که در تعریف زیر آمده‌است.

تعریف ۹.۲ **پیچیدگی نمونه ای**

تابع $m_H : (۰, ۱)^۲ \mapsto \mathbb{N}$ را پیچیدگی نمونه‌ای کلاس مفاهیم H می‌گویند. به بیانی دیگر، به تعداد نمونه‌های آموزشی مورد نیاز برای یک یادگیری موفق پیچیدگی نمونه‌ای می‌گویند.

در ادامه این بخش چند نمونه از مسایلی که قابلیت یادگیری احتمالا تقریبا درست را دارند بررسی میشوند. نخست مجموعه توابع آستانه یک بعدی [۲] که نسخه انتزاعی بسیاری از دسته بندهای واقعی است بیان میشود.

مثال ۷.۲ (تابع آستانه یک بعدی).
در مجموعه توابع آستانه یک بعدی، فرض کنید $X = \mathbb{R}$ و کلاس مفاهیم $\mathcal{C}$ مجموعه ای مفاهیم c بصورت زیر باشد

$$c(x) = \begin{cases} ۱ & x \geq \theta \\ ۰ & x < \theta \end{cases} \tag{۱۱.۲}$$

[2] One dimensional threshold function

برای هر نقطه x_0، تابع آستانه ناحیه‌ای از $\mathbb{R}$ را که در بازه $[x_0, \infty]$ باشد را برچسب مثبت زده و نقاطی که در بازه $[-\infty, x_0)$ باشند را برچسب منفی می‌زند. یا به عبارتی دیگر فرضیه یادگرفته شده به صورت زیر نمونه ها را دسته بندی می‌کند.

$$h(x) = \begin{cases} 1 & x \geq x_0 \\ 0 & x < x_0 \end{cases} \tag{۱۲.۲}$$

نقطه x_0 هر نقطه‌ای می‌تواند باشد که نقاط مثبت و منفی را از هم جدا نماید. فرضیه h را به گونه‌ای انتخاب می‌کنیم که با داده‌های آموزشی سازگار باشد. برای این منظور از الگوریتم یادگیری ۱ استفاده می‌کنیم که فرضیه h را به صورت زیر انتخاب می‌کند. در این فرضیه x_0 برابر سمت چپ ترین نمونه مثبت است..

الگوریتم ۱ الگوریتم یادگیری تابع آستانه یک بعدی

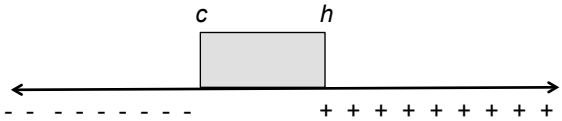
```
procedure OTF(S)
    x_0 ← ∞
    for i ← 1 to m do
        if (c(x_i) = 1) & (x_i < x_0) then
            x_0 ← x_i
        end if
    end for
    return h = x_0
end procedure
```

حال نشان میدهیم که این کلاس مفاهیم قابلیت یادگیری احتمالا تقریبا درست را دارد.

قضیه ۲.۲ (قابلیت یادگیری احتمالا تقریبا درست توابع آستانه یک بعدی).
الگوریتم ۱، قابلیت یادگیری احتمالا تقریبا درست را دارد و پیچیدیگی نمونه‌ای آن برابر است با

$$m \geq \frac{1}{\epsilon} \ln \frac{1}{\delta}. \tag{۱۳.۲}$$

برهان برای اثبات، نخست حالت کلی تر را در نظر می‌گیریم و سپس برای الگوریتم یاد شده قضیه را اثبات می‌نماییم. برای اثبات شکل زیر را در نظر بگیرید.



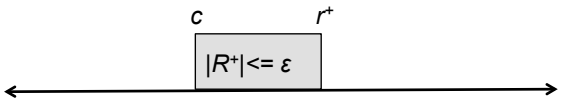
شکل ۱.۲ ناحیه خطا برای تابع آستانه یک بعدی

دراین حالت یک ناحیه مانند شکل ۱.۲ مانند شکل بالا بین c و h باقی می‌ماند که فرضیه h نقاط درون این ناحیه را به اشتباه برچسب می‌زند. هدف این است که اندازه این ناحیه کوچکتر یا مساوی ϵ باشد. اندازه به معنای طول بین c و h نیست بلکه جرم تابع توزیع در این بازه است. ما می‌خواهیم تضمین نماییم که برای هر توزیع $\mathcal{D}$ و همه نقاط ممکن، احتمال قرار گرفتن نقاط در این بازه بیش از ϵ نباشد که دو حالت زیر پیش می‌آید.

۱. فرضیه h در فاصله‌ای بیشتر از ϵ در سمت راست c قرار گیرد به این حالت B^+ می‌گوییم.

۲. فرضیه h در فاصله‌ای بیشتر از ϵ در سمت چپ c قرار گیرد به این حالت B^- می‌گوییم.

نخست حالت B^+ را در نظر بگیرید. بر اساس شکل زیر فرض کنید R^+ مشخص کننده کوچکترین ناحیه‌ای باشد که c در سمت چپ آن و جرم احتمال در آن دست کم برابر ϵ باشد یعنی $R^+ = [c, r^+]$ که $r^+ = sup\{r \geq c \mid \mathbb{P}[c, r] < \epsilon\}$. با توجه به الگوریتم ارایه شده، اگر همه نمونه‌ها در خارج از بازه R^+ قرار داشته باشند h در سمت راست ناحیه r^+ قرار می‌گیرد و هیچ نمونه آموزشی در بازه R^+ قرار نمی‌گیرید.



شکل ۲.۲ تعیین کران برای ناحیه خطا برای تابع آستانه یک بعدی

در حالتی که $\mathbb{P}[x \in R^+] < \epsilon$ باشد، مساله حل است و اگر بازه R^+ دارای اندازه دست کم ϵ باشد داریم $\mathbb{P}[x \in R^+] \geq \epsilon$ و به طور معادل رابطه $\mathbb{P}[x \notin R^+] \leq ۱ - \epsilon$ نتیجه می‌شود. برای m نمونه آموزشی می‌توان نوشت

$$\mathbb{P}[B^+] = \mathbb{P}[(x_۱ \notin \mathbb{R}^+) \wedge \ldots \wedge (x_m \notin \mathbb{R}^+] \quad (۱۴.۲)$$

$$= \mathbb{P}[x_۱ \notin \mathbb{R}^+] \ldots \mathbb{P}[x_m \notin \mathbb{R}^+] \quad (۱۵.۲)$$

$$\leq (۱ - \epsilon)^m. \quad (۱۶.۲)$$

خط دوم رابطه بالا بر این اساس نوشته شده‌است که نمونه‌ها مستقل از هم و با توزیع یکسان نمونه برداری شده‌اند. تحلیل بالا را می‌توان در باره ناحیه B^- نیز نوشت . بنابراین احتمال اینکه خطایی بیشتر از ϵ داشته باشیم را می‌توانیم به صورت زیر بدست آوریم.

$$\mathbb{P}[\mathbf{R}(h) > \epsilon] \leq \mathbb{P}[B^+ \cup B^-] \quad (۱۷.۲)$$

$$\leq \mathbb{P}[B^+] + \mathbb{P}[B^-] \quad (۱۸.۲)$$

$$\leq ۲(۱ - \epsilon)^m \quad (۱۹.۲)$$

$$\leq ۲e^{-\epsilon m}. \quad (۲۰.۲)$$

خط دوم رابطه بالا براساس کران اجتماعات و خط سوم بر اساس تقارن نوشته شده‌اند و خط آخر بر اساس نابرابری $۱ + x \leq e^x$ نوشته شده است. هدف این است که این احتمال از δ کوچکتر باشد و یا به عبارتی دیگر داشته باشیم

$$\mathbb{P}[\mathbf{R}(h) > \epsilon] \leq ۲e^{-\epsilon m} \leq \delta. \quad (۲۱.۲)$$

حال اگر نابرابری بالا را بر اساس m حل نماییم خواهیم داشت

$$m \geq \frac{۱}{\epsilon} \ln \frac{۲}{\delta}. \quad (۲۲.۲)$$

■ با توجه به الگوریتم یاد شده، هیچ گاه حالت B^- رخ نمی دهد و بنابراین ضریب ۲ در رابطه بالا حذف شده و قضیه اثبات می‌گردد.

در عمل مجموعه آموزشی که تعداد نمونه ها را مشخص می‌نماید در اختیار الگوریتم قرار داده می‌شود و هدف این است که کران خطای واقعی را تخمین بزنیم. با استفاده از قضیه بالا و با محاسبات ساده به سادگی می‌توان نشان داد که با احتمال دست کم $1-\delta$ نابرابری زیر برقرار است.

$$\mathbf{R}(h) \leq \frac{1}{m} \ln \frac{2}{\delta}. \tag{۲۳.۲}$$

برای بدست آوردن معادله بالا، کافی است معادلعه (۲۲.۲) را براساس ϵ بازنویسی نماییم که به نتیجه زیر می‌رسیم.

$$\epsilon \geq \frac{1}{m} \ln \frac{2}{\delta}. \tag{۲۴.۲}$$

و از تعریف یادگیری احتمالا تقریبا درست، نابرابری $\mathbf{R}(h) \leq \epsilon$ را داریم. لذا کوچکترین مقداری که برای ϵ می‌توان متصور شد برابر است با $\frac{1}{m} \ln \frac{2}{\delta}$ و در نتیجه خواهیم داشت

$$\mathbf{R}(h) \leq \frac{1}{m} \ln \frac{2}{\delta}. \tag{۲۵.۲}$$

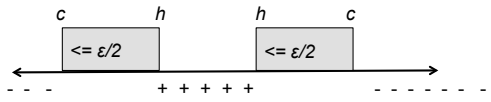
حال کلاس فرضیه را از توابع آستانه یک بعدی به بازه گسترش می‌دهیم.

مثال ۸.۲ (یادگیری بازه).
در مجموعه توابع بازه، فرض کنید $X = \mathbb{R}$ و کلاس مفاهیم $\mathcal{C}$ مجموعه‌ای مفاهیم c بصورت زیر باشد

$$c(x) = \begin{cases} 1 & x \in [a,b] \\ 0 & x \notin [a,b] \end{cases} \tag{۲۶.۲}$$

این مساله را با کمک توابع آستانه یک بعدی حل می‌کنیم. در این مساله دو ناحیه مرزی خطا مطابق شکل ۳.۲ داریم و می‌خواهیم که اندازه هرکدام حداکثر $\frac{\epsilon}{2}$ باشد و بنابراین چهار حالت مختلف داریم. اگر از پاسخ مساله قبل استفاده کنیم خواهیم داشت

$$m \geq \frac{2}{\epsilon} \ln \frac{4}{\delta}. \tag{۲۷.۲}$$



شکل ۳.۲ تعیین کران برای ناحیه خطا برای تابع بازه

مثال ۹.۲ (پیچیدگی نمونه ای مستطیل های با اضلاع موازی محورها).
چند نمونه برای یادگیری موفق مساله مستطیل های با اضلاع موازی محورها مورد نیاز است؟
در این مثال، فضای فرضیه ها را مستطیل‌هایی که اضلاع آنها موازی محورهای مختصات باشد در نظر می‌گریم و الگوریتم یادگیری تلاش می‌کند که کوچکترین مستطیل سازگار با داده های آموزشی را پیدا کند. بنابراین می‌بایست $\hat{\mathbf{R}}(h) = 0$. فرض کنید c مفهوم مستطیل باشد همانگونه که در شکل ۴.۲ نشان داده شده است فرضیه تولید شده که با h نشان داده می‌شود درون مستطیل c قرار دارد. نوارهای بین مستطیل های c و h خطای واقعی فرضیه تولید شده h را نشان می‌دهند. هدف پیدا کردن تعداد نمونه های لازم است به گونه ای که احتمال قرار گرفتن یک نمونه مثبت در این چهار نوار کوچکتر یا مساوی ϵ باشد.اگر برای هر کدام از این

نوارها بتوانیم تضمین نماییم که کران بالای این احتمال $\frac{\epsilon}{۴}$ است لذا مجموع این احتمال ها حداکثر برابر خواهد بود با $۴(\frac{\epsilon}{۴}) = \epsilon$.
دقت شود که چهار گوشه نوارها را دو بار محاسبه نموده ایم لذا خطا از $۴(\frac{\epsilon}{۴}) = \epsilon$ کمتر خواهد بود.
احتمال اینکه یک نمونه تصادفی در یک نوار نباشد برابر است با $۱ - \frac{\epsilon}{۴}$ و احتمال اینکه m نمونه در این نوار نباشند برابر است با $(۱ - \frac{\epsilon}{۴})^m$ و احتمال اینکه این m نمونه در هیچکدام از این چهار نوار نباشند حداکثر برابر است با $۴(۱ - \frac{\epsilon}{۴})^m$ که میبایست حداکثر برابر باشد با δ. بنابراین داریم.

$$\begin{aligned} ۴\left(۱ - \frac{\epsilon}{۴}\right)^m &\leq \delta \\ \left(۱ - \frac{\epsilon}{۴}\right)^m &\leq \frac{\delta}{۴} \end{aligned}$$

با استفاده از نابرابری $۱ - x \leq e^{-x}$ خواهیم داشت.

$$\left(۱ - \frac{\epsilon}{۴}\right)^m \leq \exp\left[-\frac{\epsilon m}{۴}\right] \leq \frac{\delta}{۴}$$

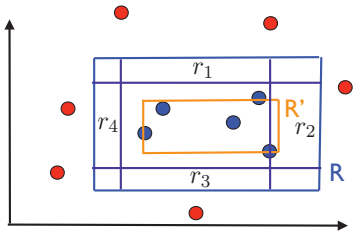
اگر از طرفین نابرابری بالا لگاریتم گرفته شود خواهیم داشت

$$-\frac{\epsilon m}{۴} \leq \log\frac{\delta}{۴}$$

بنابراین کمترین تعداد نمونه های مورد نیاز برای یادگیری موفق این مساله برابر است با

$$m \geq \frac{۴}{\epsilon}\log\frac{۴}{\delta}$$

در نتیجه مساله یادگیری مستطیل های با اضلاع موازی محورها قابلیت یادگیری احتمالا تقریبا درست را دارا است.



شکل ۴.۲ خطای بین فرضیه تولید شده و مفهوم

۱.۳.۲ پیچیدگی نمونه ای برای فضای فرضیه متناهی

در مثال های یاد شده فرضیه ای که توسط الگوریتم تولید می‌شود با داده‌های اموزشی سازگار بودند. در این بخش کران پیچیدگی نمونه‌ای و به صورت معادل کران خطای واقعی را برای فرضیه های سازگار و حالتی که $|H|$ متناهی باشد محاسبه می‌کنیم. چون هدف پیدا کردن فرضیه سازگار با داده‌های آموزشی است لذا فرض می‌کنیم که فضای فرضیه H حاوی مفهوم c باشد یا به عبارتی دیگر شرط $c \in H$ برقرار باشد. قضیه زیر این کران را مستقل از الگوریتم یادگیری محاسبه می‌نماید.

قضیه ۳.۲ (پیچیدگی نمونه‌ای فضای فرضیه متناهی).
اگر H یک مجموعه متناهی از فرضیه ها و مجموعه $S = \{(x_۱, y_۱), (x_۲, y_۲), \ldots, (x_m, y_m)\}$ مجموعه آموزشی باشد که از توزیع $\mathcal{D}$ و بطور مستقل از هم نمونه برداری شده باشند و توسط مفهوم $c \in \mathcal{C}$ برچسب گذاری شده باشند. آنگاه برای هر $۰ < \epsilon < \frac{۱}{۲}$ و $۰ < \delta < \frac{۱}{۲}$، اگر الگوریتم یادگیری A فرضیه سازگار $h \in H$ را تولید نماید نابرابری‌های زیر با احتمال دست‌کم $۱ - \delta$ برقرار هستند..

$$m \geq \frac{۱}{\epsilon}\left[\log|H| + \log\frac{۱}{\delta}\right] \tag{۲۸.۲}$$

$$\mathbf{R}(h) \leq \frac{۱}{m}\left[\log|H| + \log\frac{۱}{\delta}\right] \tag{۲۹.۲}$$

در رابطه‌های بالا، $\log|H|$ را می‌توان اندازه بازنمایی فرضیه های فضای H یعنی تعداد بیت های لازم برای بازنمایی فرضیه های این فضا دانست. در واقع این معیار نشان دهنده غنای فضای فرضیه است به گونه‌ای که هرچه این فضا بزرگتر باشد احتمالا انعطاف‌پذیرتر است و می‌تواند فرضیه نزدیک به مفهوم را در خود داشته باشد. در ادامه نشان می‌دهیم که که این معیار الزما به‌خوبی غنای فضای فرضیه را نشان نمی‌دهد.

برهان نخست نابرابری را (۲۸.۲) بدست میآوریم و سپس نابرابری (۲۹.۲) را. احتمال اینکه فرضیه سازگاری وجود داشته باشد که خطای واقعی آن بیشتر از ϵ برابر است با

$$\begin{aligned}\mathbb{P}\left[\exists h \in H \;\middle|\; \hat{\mathbf{R}}(h) = ۰ \& \mathbf{R}(h) > \epsilon\right] &= \mathbb{P}\left[(\hat{\mathbf{R}}(h_۱) = ۰ \wedge \mathbf{R}(h_۱) > \epsilon) \vee (\hat{\mathbf{R}}(h_۲) = ۰ \wedge \mathbf{R}(h_۲) > \epsilon) \vee \ldots\right] \\ &\leq \sum_{h_i \in H} \mathbb{P}\left[\hat{\mathbf{R}}(h_i) = ۰ \wedge \mathbf{R}(h_i) > \epsilon\right] \qquad (۳۰.۲) \\ &\leq \sum_{h_i \in H} \mathbb{P}\left[\hat{\mathbf{R}}(h_i) = ۰ \;\middle|\; \mathbf{R}(h_i) > \epsilon\right]. \qquad (۳۱.۲)\end{aligned}$$

نابرابری (۳۰.۲) با استفاده از نابرابری کران اجتماعات بدست آمده است

$$\mathbb{P}\left[\bigcup_{i=۱}^{K} A_i\right] \leq \sum_{i=۱}^{K} \mathbb{P}\left[A_i\right] \tag{۳۲.۲}$$

و نابرابری (۳۱.۲) نیز با استفاده از قاعده احتمال شرطی بدست آمده است. حال هر $h \in H$ را در نظر بگیرید که $\mathbf{R}(h) > \epsilon$ باشد انگاه احتمال اینکه h با داده‌های آموزشی سازگار باشد را می‌توان به صورت زیر محاسبه نمود.

$$\mathbb{P}\left[\hat{\mathbf{R}}(h) = ۰ | \mathbf{R}(h) > \epsilon\right] \leq (۱ - \epsilon)^m \tag{۳۳.۲}$$

نابرابری بالا از روش زیر بدست آمده است. فرض کنید که $\mathbf{R}(h) > \epsilon$ باشد. احتمال اینکه h یک نمونه را اشتباه دسته بندی نماید بزرگتر از ϵ است و در نتیجه احتمال اینکه h یک نمونه را درست دسته بندی نماید کوچکتر یا مساوی $۱ - \epsilon$ خواهد بود و احتمال اینکه h همه m نمونه را درست دسته بندی نماید کوچکتر یا مساوی $(۱ - \epsilon)^m$ است. باقرار دادن رابطه بالا در نابرابری (۳۱.۲) خواهیم داشت.

$$\sum_{h_i \in H} \mathbb{P}\left[\hat{\mathbf{R}}(h_i) = ۰ | \mathbf{R}(h_i) > \epsilon\right] \leq |H|(۱ - \epsilon)^m \leq \delta \tag{۳۴.۲}$$

با استفاده از نابرابری $۱ - x \leq e^{-x}$ و نابرابری بالا خواهیم داشت.

$$|H|(۱ - \epsilon)^m \leq |H| \exp(-\epsilon m) \leq \delta \tag{۳۵.۲}$$

با ساده سازی این نابرابری، نابرابری زیر بدست میآید.

$$m \geq \frac{1}{\epsilon}\left[\log|H| + \log\frac{1}{\delta}\right]. \tag{۳۶.۲}$$

برای بدست آوردن نابرابری (۲۹.۲) کافی است نابرابری (۲۸.۲) را براساس ϵ حل نماییم، نابرابری زیر بدست می‌آید.

$$\epsilon \geq \frac{1}{m}\left[\log|H| + \log\frac{1}{\delta}\right]. \tag{۳۷.۲}$$

با توجه به اینکه کران بالای خطای واقعی ϵ است لذا همیشه خطای واقعی از کمترین مقدار ϵ کمتر خواهد بود و در نتیجه نابرابری (۲۹.۲) بدست می‌آید و قضیه اثبات می‌گردد. ■

قضیه بالا بیان می‌دارد که هنگامی که H متناهی است هر الگوریتم سازگار A یک الگوریتم یادگیری احتمالا تقریبا درست است. هزینه ای که برای الگوریتم‌های سازگار می‌پردازیم این است که از مجموعه H های بزرگتری فرضیه را انتخاب می‌کنیم که مفهوم نهایی c در آن است یعنی $c \in H$. یکی از مشکلات نابرابری (۳۵.۲) این است که اگر $|H|$ خیلی بزرگ باشد ممکن است سمت چپ این نابرابری بیشتر از یک باشد. یکی از اشکالات این کران این است که این کران برای $|H|$ های بزرگ سفت نیست. یعنی اگر $|H| = \infty$ باشد این کران اصلا مناسب نیست. حال با توجه به قضیه بالا، تعداد نمونه های لازم برای چند مساله یادگیر را بدست می‌آوریم.

مثال ۱۰.۲ (یادگیری ترکیب عطفی از n متغیر).
فرض کنید H فضای فرضیه‌ای باشد که از ترکیب عطفی چند متغیر(ویژگی) ساخته شود. برای یادگیری این مساله به چند نمونه نیاز است؟
در این مساله به جای هر متغیر می‌توان آن متغیر، نقیض آن و یا ? را قرار داد در نتیجه $|H| = ۳^n$ خواهد بود که n تعداد ویژگی‌های مساله است. بنابراین با استفاده از نابرابری (۲۸.۲) داریم

$$\begin{aligned} m &\geq \frac{1}{\epsilon}\left[\log|H| + \log\frac{1}{\delta}\right] \\ &= \frac{1}{\epsilon}\left[\log ۳^n + \log\frac{1}{\delta}\right] \\ &= \frac{1}{\epsilon}\left[n\log ۳ + \log\frac{1}{\delta}\right]. \end{aligned}$$

چون پیچیدگی نمونه‌ای برای یادگیری این مساله یک تابع چندجمله‌ای از پارامترهایش است پس این مساله قابلیت یادگیری احتمالا تقریبا درست را دارد.

در ادامه مسال یادگیری تابع دودویی را بررسی می‌کنیم.

مثال ۱۱.۲ (یادگیری از فضای همه توابع دودویی با n متغیر).
فرض کنید H فضای فرضیه همه توابع دودویی باشد که از n متغیر ساخته شود. برای یادگیری این مساله به چند نمونه نیاز است؟
در این مساله فضای فرضیه دارای $|H| = ۲^{۲^n}$ فرضیه خواهد بود. بنابراین با استفاده از نابرابری (۲۸.۲) داریم

$$\begin{aligned} m &\geq \frac{1}{\epsilon}\left[\log|H| + \log\frac{1}{\delta}\right] \\ &= \frac{1}{\epsilon}\left[\log ۲^{۲^n} + \log\frac{1}{\delta}\right] \\ &= \frac{1}{\epsilon}\left[۲^n\log ۲ + \log\frac{1}{\delta}\right]. \end{aligned}$$

با توجه به اینکه پیچیدگی نمونه ای چند جمله ای بر اساس n نیست لذا قابلیت یادگیری احتمالا تقریبا درست را ندارد.

اگر n را یک عدد ثابت و کوچک در نظر گیریم میتوانیم ادعا کنیم که این کلاس مفاهیم قابلیت یادگیری احتمالا تقریبا درست را دارد اما همچنان پیچیدیگی نمونه‌ای با n به صورت نمایی رشد می‌کند که با تعریف یادگیری احتمالا تقریبا درست در تناقض است. برای برخی از توابع دودویی، تعداد عبارت های داخل فرضیه 2^n است که سبب نمایی شدن پیچیدگی نمونه‌ای می‌شود. اما برای بسیاری از مسایل ساده‌تر، بازنمایی بسیار ساده تر است و نیاز به پیچیدگی نمونه‌ای نمایی ندارد. آز آنجایی که میبایست تعداد نمونه ها برای تمام مفاهیم $c \in \mathcal{C}$ یک تابع چندجمله‌ای براساس پارامترهایش باشد کران قضیه برای مثال بالا نتیجه نمی‌دهد که این مساله قابلیت یادگیری احتمالا تقریبا درست را دارد اما این قضیه بیان نمی کند که این مساله قابلیت یادگیری احتمالا تقریبا درست را ندارد. زیرا کران مشخص شده در قضیه، یک کران سفت نیست و ما نمی‌توانیم از این قضیه برای نشان دادن قابلیت یادگیری احتمالا تقریبا درست این مساله استفاده کنیم. پرسش اینکه آیا این مساله قابلیت یادگیری احتمالا تقریبا درست را دارد یا نه برای بیش از دو دهه، یک مساله باز است.

قضیه ۴.۲ (قابلیت یادگیری احتمالا تقریبا درست فضای فرضیه متناهی).
هر فضای فرضیه متناهی بالقوه قابلیت یادگیری احتمالا تقریبا درست را دارد.

۴.۲ مدل یادگیری احتمالا تقریبا درست بدون پیش فرض

پیش از این فرض کردیم که $\mathcal{C} \subset H$ است و الگوریتم یادگیری یک فرضیه سازگار را تولید میکند. براین اساس نابرابری (۳۵.۲) مهم است زیرا تعداد نمونه های کافی برای یک یادگیری موفق در حالتی که خطای آموزشی صفر است را مشخص مینماید. در این نابرابری فرض شده است که $\hat{\mathbf{R}}(h) = 0$ باشد. اما در بیشتر کاربردها به دلایل زیر ممکن است فرضیه تولید شده سازگار نباشد و در نتیجه خطای آموزشی بیش از صفر داریم. الف) فضای فرضیه H به اندازه کافی غنی نیست که توانایی نشان دادن مفهوم c را داشته باشد یا به عبارتی دیگر $c \notin H$. ب) از نظر محاسباتی پیدا کردن پیدا کردن فرضیه h سازگار امکان پذیر نباشد و ج) ممکن است c قطعی و جود نداشته باشد و یک تابع تصادفی باشد. در صورتیکه H شامل c نباشد ممکن است خطای آموزش صفر نباشد. در نتیجه در این مسایل، خطای آموزشی بیشتر از صفر داریم. البته فرضیه بدست آمده در صورتی مناسب است که دارای خطای آموزش و واقعی کم باشد. علاوه بر مشکل یاد شده، ممکن است داده ها نیز نویزی باشند و نتوان یک فرضیه سازگار با داده های آموزشی پیدا کرد. لذا در این حالت نیز خطای آموزشی بزرگتر از صفر داریم. با توجه به دلایلی که گفته شد ما نیاز داریم تا تعاریف و فرضیات مدل های پیشین را بازنگری کنیم. پیش از این فرض کردیم که نمونه ها از توزیع $\mathcal{D}$ نمونه برداری می‌شوند و سپس با تابع معین c که نقش آموزگار را دارد برچسب زده می‌شوند. بنابراین برچسب ها به صورت احتمالاتی مشخص نمی‌شوند. در این جا فرض می‌کنیم که یک تابع توزیع احتمال $\mathcal{D}$ روی زوج نمونه‌های (x, y) وجود دارد که نمونه ها از این توزیع نمونه برداری می‌شوند. بنابراین خطای واقعی به صورت زیر اصلاح می‌شود.

$$\begin{aligned} \mathbf{R}(h) &= \mathop{\mathbb{P}}_{(x,y)\sim\mathcal{D}}[h(x) \neq y] \\ &= \sum_{h(x)\neq y} \mathcal{D}(x, y). \end{aligned} \tag{۳۸.۲}$$

هدف این است که فرضیه hای پیدا کنیم که با احتمال بالا دارای خطای واقعی پایینی باشد. اما در شرایط بیان شده ممکن است چنین فرضیه‌ای وجود نداشته باشد و ما می‌توانیم فرضیه‌ای را پیدا کنیم که نزدیک به بهترین فرضیه درون H باشد. در نتیجه هدف پیدا کردن فرضیه hای

است که خطای واقعی آن نزدیک به $\min_h \mathbf{R}(h)$ باشد. همچنین مدل های پیشین فرض می‌کردند که الگوریتمی وجود دارد که توانایی پیدا کردن فرضیه سازگار با داده‌های آموزشی را دارد. این فرض در این حالت دیگر وجود ندارد. بنابراین الگوریتم‌هایی را بررسی میکنیم که فرضیه $h \in H$ را پیدا می‌کنند که بیشترین سازگاری را با داده‌ها داشته باشد. این سازگاری بر اساس خطای تجربی روی داده‌های آموزشی تعریف می‌شود. یک روش این است که در این مسایل ما به دنبال فرضیه‌ای باشیم که خطای آموزش را کمینه نماید. بیشتر الگوریتم‌های یادگیری ماشین که هم اکنون به کار می‌روند از این گونه هستند. به این الگوریتم ها، الگوریتم‌های کمینه سازی خطای تجربی [۳] می‌گویند. دلیل به کارگیری این الگوریتم ها این است که کمینه سازی خطای تجربی تقریب خوبی به کمینه سازی خطای واقعی است. دلایل دیگر برای بکارگیری این الگوریتم ها در یادگیری ماشین را در بخش‌های بعدی بیان خواهیم نمود. در واقع الگوریتم‌های کمینه‌سازی خطای تجربی فرضیه h_{erm} را برای مجموعه آموزشی S به صورت زیر انتخاب میکنند.

$$h_{erm} = \operatorname*{argmin}_{h \in H} \hat{\mathbf{R}}(h). \tag{۳۹.۲}$$

در الگوریتم‌های کمینه سازی خطای تجربی، دو نکته وجود دارد که باید به درستی انتخاب شوند. نخست آنکه فضای فرضیه H به درستی انتخاب شود و دوم آنکه شیوه کمینه سازی خطای تجربی که بر اساس تابع هزینه (تابع خطا) تعیین می‌گردد به درستی انتخاب گردد. حال برای روشن شدن موضوع یک مثال بیان می‌شود.

مثال ۱۲.۲ (رگرسیون خطی).
مساله رگرسیون خطی را در نظر بگیرید که مجموعه آموزشی به صورت $S = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)\}$ که $\mathbf{x}_i \in \mathbb{R}^n$ و $y_i \in \mathbb{R}$ و نمونه‌های مجموعه آموزشی براساس توزیع $\mathcal{D}$ و به صورت مستقل نمونه برداری شده باشند. همچنین فرض کنید که $\mathbf{A}$ ماتریسی باشد که سطرهایش مقادیر $\mathbf{x}_i$ ها و مجموعه فرضیه‌ها برابر $H = \{h_w : \mathcal{X} \mapsto \mathcal{Y} | h_w(x) = \langle \mathbf{x}, \mathbf{w} \rangle\}$ باشد. اگر تابع خطای تجربی (تابع هزینه) را به صورت زیر تعریف کنیم

$$\hat{\mathbf{R}}(h) = \frac{1}{m} \sum_{i=1}^{m} [h(\mathbf{x}_i) - y_i]^2, \tag{۴۰.۲}$$

یک الگوریتم کمینه سازی خطای تجربی، فرضیه زیر را تولید میکند.

$$\mathbf{w}^* = \operatorname*{argmin}_{\mathbf{w} \in \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^{m} [h(\mathbf{x}_i) - y_i]^2 \tag{۴۱.۲}$$

$$= \operatorname*{argmin}_{w \in \mathbb{R}^n} \|\mathbf{A}\mathbf{w} - \mathbf{y}\|^2 \tag{۴۲.۲}$$

که $\mathbf{y}$ برداری است با مقادیر y_i ها پر شده است. با حل این معادله فرضیه زیر بدست میآید.

$$\mathbf{w}^* = \left(\mathbf{A}^T \mathbf{A}\right)^{-1} \mathbf{A}^T \mathbf{y}. \tag{۴۳.۲}$$

در واقع در این مساله، الگوریتم کمینه سازی خطای تجربی یک پاسخ بسته برای ما تولید مینماید اما همیشه چنین نیست.

در الگوریتم‌های کمینه سازی خطای تجربی یک پرسش مهم وجود دارد که بهترین فرضیه یعنی فرضیه‌ای که کمترین خطای واقعی را داشته باشد کدام است و مقدار خطای واقعی این فرضیه چه اندازه است؟ بهترین فرضیه را به صورت h_{opt} و خطای واقعی آن را به صورت

[3] Empirical risk minimization algorithms (ERM)

$\mathbf{R}^*$ نشان می‌دهیم و برابر است با

$$\mathbf{R}^* = \inf_{h \in H} \mathbf{R}(h). \tag{۴۴.۲}$$

بهترین مقدار خطای واقعی که به خطای بیز معروف است با دسته بند زیر که به دسته‌بند بهینه‌بیز معروف است قابل حصول است

$$h_{opt} = \begin{cases} ۱ & if\ \mathbb{P}[y = ۱ \mid x] \geq \frac{۱}{۲} \\ ۰ & if\ \mathbb{P}[y = ۱ \mid x] < \frac{۱}{۲} \end{cases} \tag{۴۵.۲}$$

با توجه به اینکه یادگیر به توزیع $\mathcal{D}$ دسترسی ندارد و آن را نمی شناسد برای دسته بند بهینه بیز نمی‌توان مقدار خطای بیز را محاسبه نمود و بر اساس قضیه زیر می‌توان نشان داد که هیچ دسته بندی خطای واقعی کمتری نسبت به دسته بند بهینه بیز تولید نمی‌کند.

قضیه ۵.۲ (خطای دسته بند بهینه بیز).
خطای دسته بند بهینه بیز برابر $\mathbf{R}^*$ است. به عبارتی دیگر هیچ دسته بندی وجود ندارد که خطایی کمتر از خطای دسته بند بهینه بیز را تولید نماید.

برهان اثبات این قضیه را به عنوان تمرین انجام دهید. ■

هدف همه روش های یادگیری ماشین این است که $\min_{h \in H} \mathbf{R}(h)$ را پیدا نمایند. برای این منظور برای مجموعه آموزشی S داده شده، خطای تجربی را کمینه می‌کنیم. برای اینکه بتوان از این رویکرد استفاده نمود باید بتوانیم نشان دهیم که کمینه‌سازی حطای تجربی سبب کمینه‌سازی خطای واقعی می‌شود یا دست کم می‌توان کران خطای واقعی را بر اساس خطای تجربی پیش‌بینی نمود. به بیانی دیگر باید بگوییم که برای همه فرضیه‌های $h \in H$ با احتمال $۱ - \delta$ نابرابری زیر برقرار است.

$$\left|\mathbf{R}(h) - \hat{\mathbf{R}}(h)\right| \leq \epsilon \tag{۴۶.۲}$$

یا به بیانی دیگر کران پایین m را باید به گونه‌ای پیدا کنیم که برای همه $\epsilon, \delta \in (۰, ۱)$ نابرابری زیر برقرار باشد.

$$\mathbb{P}\left[\left|\mathbf{R}(h_{erm}) - \min_{h \in H} \mathbf{R}(h)\right| > \epsilon\right] < \delta \tag{۴۷.۲}$$

که h_{erm} فرضیه‌ای است که الگوریتم کمینه سازی خطای تجربی آن را تولید می‌نماید.نابرابری (۴۶.۲) این نتیجه را می‌دهد که اگر h_{erm} خطای تجربی را کمینه نماید آنگاه برای همه $h \in H$ها داریم

$$\begin{aligned} \mathbf{R}(h_{erm}) &\leq \hat{\mathbf{R}}(h_{erm}) + \epsilon && (۴۸.۲) \\ &\leq \hat{\mathbf{R}}(h) + \epsilon && (۴۹.۲) \\ &\leq (\mathbf{R}(h) + \epsilon + \epsilon && (۵۰.۲) \\ &= \mathbf{R}(h) + ۲\epsilon. && (۵۱.۲) \end{aligned}$$

بدلیل اینکه h_{erm} خطای تجربی را کمینه می‌کند رابطه (۴۹.۲) بدست آمده‌است. رابطه (۵۰.۲) به صورت زیر بدست آمده‌است. چون h_{erm} خطای تجربی را کمینه می‌نماید از نابرابری (۴۶.۲) داریم

$$-\epsilon \leq \mathbf{R}(h) - \hat{\mathbf{R}}(h) \leq \epsilon \tag{۵۲.۲}$$

با ساده سازی نابرابری بالا به نابرابری زیر می‌رسیم.

$$\hat{\mathbf{R}}(h) \leq \mathbf{R}(h) + \epsilon. \tag{۵۳.۲}$$

نابرابری بالا بیان می‌دارد که خطای واقعی فرضیه‌ای که از کمینه سازی خطای تجربی بدست می‌آید از کمینه خطای همه فرضیه‌ها بعلاوه ۲ϵ بیشتر نخواهدبود. در نتیجه این فرضیه خطای واقعی نزدیک به کران پایین خطا برای همه فرضیه‌های مجموعه H را تولید می‌نماید برای ثابت نمودن این نابرابری، می‌بایست همگرایی یکنواخت را برای همه $h \in H$ ثابت نماییم یعنی نابرابری (۴۶.۲) را اثبات نماییم. این نابرابری بیان می‌کند که الگوریتم کمینه‌سازی خطا یک یادگیر احتمالا تقریبا درست بدون پیش فرض است. پیش از اینکه همگرایی یکنواخت را برای همه $h \in H$ ثابت نماییم، نخست مدل توسعه یافته مدل یادگیری احتمالا تقریبا درست را برای حالتی که فرضیه بدست آمده سازگار نباشد بیان می‌کنیم. این مدل یادگیری را مدل یادگیری بدون پیش فرض [۴] می‌گویند و به صورت زیر تعریف می‌شود زیرا فرضیاتی که پیش از این داشتیم را دیگر ندارد.این یادگیری که فرض نمی‌کند که $c \notin H$ و تنها فرضیه‌ای را پیدا می‌کند که خطای آموزش را کمینه نماید یادگیر بدون پیش فرض نامیده می‌شود زیرا الزامی در باره شرط $C \subset H$ ندارد.

تعریف ۱۰.۲ **یادگیری احتمالا تقریبا درست بدون پیش فرض** فرض کنید H یک فضای فرضیه باشد. الگوریتم یادگیری A یک الگوریتم یادگیری احتمالا تقریبا درست بدون پیش فرض است اگر یک چند جمله‌ای p وجود داشته‌باشد به گونه‌ای که برای هر $\epsilon \in (۰, ۱)$ و $\delta \in (۰, ۱)$ و برای هر توزیع $\mathcal{D}$ روی $\mathcal{X} \times \mathcal{Y}$ برای هر مجموعه آموزشی با اندازه

$$m \geq p\left(\frac{۱}{\epsilon}, \frac{۱}{\delta}, |\mathbf{x}|, |C|\right) \tag{۵۴.۲}$$

نابرابری زیر برای فرضیه h_s که توسط الگوریتم یادگیری A با دیدن مجموعه آموزشی S تولید می‌شود برقرار است.

$$\mathbb{P}\left[\left|\mathbf{R}(h_s) - \min_{\hat{h} \in H} \mathbf{R}(\hat{h})\right| \leq \epsilon\right] \geq ۱ - \delta \tag{۵۵.۲}$$

براساس تعریف بالا، می‌توان الگوریتم کارا را در این مدل به صورت زیر تعریف نمود.

تعریف ۱۱.۲ **الگوریتم یادگیری احتمالا تقریبا درست بدون پیش فرض کارا** اگر الگوریتم A که در تعریف ۱۰.۲ آمده است در زمان چندجمله‌ای $p(\frac{۱}{\epsilon}\frac{۱}{\delta}, |x|, |c|)$ اجرا شود انگاه به این الگوریتم، یک الگوریتم یادگیری احتمالا تقریبا درست بدون پیش فرض کارا می‌گویند.

همان‌گونه که بیان شد الگوریتم کمینه‌سازی خطا یک الگوریتم یادگیر احتمالا تقریبا درست بدون پیش‌فرض است. برای اینکه این ویژگی را نشان بدهیم باید اثبات کنیم که برای همه فرضیه‌های $h \in H$، با احتمال دستکم $۱ - \delta$ روی یک مجموعه‌ای که به‌صورت تصادفی از فضای نمونه انتخاب شده‌باشد نابرابری (۵۵.۲) برقرار است. این نیازمندی‌ها توسط ویژگی همگرایی یکنواخت بیان می‌شود که در ادامه بیان می‌شود.

[4] Agnostic learning model

تعریف ۱۲.۲ **همگرایی یکنواخت** فضای ویژگی H با توجه به دامنه $\mathcal{X} \times \mathcal{Y}$ و تابع خطا (هزینه یا زیان) دارای ویژگی همگرایی یکنواخت است اگر یک تابع $m_H^{UC} : (۰,۱)^۲ \mapsto \mathbb{N}$ وجود داشته‌باشد به‌گونه‌ای که برای هر $\epsilon \in (۰,۱)$ و $\delta \in (۰,۱)$ و برای هر توزیع $\mathcal{D}$ روی $\mathcal{X} \times \mathcal{Y}$، اگر مجموعه آموزشی S با اندازه $m \geq m_H^{UC}(\epsilon, \delta)$ به‌صورت مستقل و با توزیع $\mathcal{D}$ نمونه‌برداری شده‌باشند آنگاه نابرابری زیر برقرار است.

$$\mathbb{P}\left[\forall h \in H \middle| \mathbf{R}(h) - \hat{\mathbf{R}}(h) \middle| \leq \epsilon\right] \geq ۱ - \delta$$

واژه یکنواخت در تعریف بالا به این معنی است که داشتن یک مجموعه آموزشی با اندازه مشخص که برای همه فرضیه‌های $h \in H$ و همه توزیع های ممکن روی دامنه کار می‌کند. مشابه پیچیدگی نمونه‌ای که پیش از این داشتیم، تابع m_H^{UC} به کمترین اندازه مجموعه آموزشی اشاره می‌کند که برای آن ویژگی همگرایی یکنواخت برقرار است. در فصل‌های بعد نشان خواهیم‌داد که ویژگی یکنواخت شرط لازم و کافی برای قابلیت یادگیری است.

پیش از اینکه رابطه (۴۶.۲) را اثبات کنیم نخست حالتی ساده که تنها یک فرضیه داشته باشیم در نظر می‌گیریم و آن را بررسی می‌کنیم.

قضیه ۶.۲ (کران خطا برای یادگیر بدون پیش فرض).
اگر S یک مجموعه آموزشی با m نمونه که بصورت مستقل و با توزیع یکسان از $\mathcal{D}$ نمونه برداری شده باشد آنگاه برای هر ثابت $\delta > ۰$ و هر فرضیه ثابت $h : \mathcal{X} \mapsto \{۰,۱\}$ نابرابری زیر برقرار است

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}(h) + \sqrt{\frac{\ln \frac{۲}{\delta}}{۲m}} \tag{۵۶.۲}$$

پیش از اثبات این قضیه، نخست نابرابری هافدینگ را که در اثبات قضیه به کار می‌رود را شرح می‌دهیم. فرض کنید که $X_۱, \ldots, X_m$ متغیرهای تصادفی مستقل باشند که برای همه iها شرط $X_i \in [a_i, b_i]$ برقرار باشد. انگاه برای هر $\epsilon > ۰$، نابرابری زیر برای $S_m = \sum_{i=۱}^m X_i$ برقرار است.

$$\mathbb{P}\left[|S_m - \mathbb{E}\left[S_m\right]| \geq \epsilon\right] \leq ۲ \exp\left[-۲\frac{\epsilon^۲}{\sum_i (b_i - a_i)^۲}\right] \tag{۵۷.۲}$$

برهان اثبات این قضیه با کمک نابرابری (۵۷.۲) به صورت زیر انجام می‌شود. برای هر $i \in \{۱, ۲, \ldots, m\}$، متغیر z_i را به صورت زیر تعریف می‌کنیم.

$$z_i = \begin{cases} ۱ & if \;\; h(x_i) \neq y_i \\ ۰ & if \;\; h(x_i) = y_i \end{cases} \tag{۵۸.۲}$$

با توجه به $\hat{\mathbf{R}}(h) = \frac{۱}{m}\sum_{i=۱}^m z_i$ و $\mathbf{R}(h) = \mathbb{E}\left[\hat{\mathbf{R}}(h)\right]$. و نابرابری هافدینگ داریم

$$\mathbb{P}\left[\left|\hat{\mathbf{R}}(h) - \mathbf{R}(h)\right| \geq \epsilon\right] \leq ۲ \exp\left(-۲m\epsilon^۲\right) \leq \delta. \tag{۵۹.۲}$$

با توجه به اینکه سمت راست نابرابری بالا باید کوچکتر از δ باشد و از طرفین این نابرابری log گرفته شود خواهیم داشت.

$$\epsilon \geq \sqrt{\frac{\ln \frac{۲}{\delta}}{۲m}} \tag{۶۰.۲}$$

و با استفاده از تعریف مدل یادگیری احتمالا تقریبا درست بدون پیش فرض برای هر $\delta > ۰$، نابرابری زیر با احتمال دست کم $(۱ - \delta)$ برقرار است.

$$\left|\mathbf{R}(h) - \hat{\mathbf{R}}(h)\right| \leq \sqrt{\frac{\ln \frac{۲}{\delta}}{۲m}} \tag{۶۱.۲}$$

■

نتیجه ۱.۲ .
فرضیه $h : \mathcal{X} \mapsto \{۰, ۱\}$ را در نظر بگیرید آنگاه برای هر $\delta > ۰$، نابرابری زیر با احتمال دست کم $(۱ - \delta)$ برقرار است.

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}(h) + \sqrt{\frac{\ln \frac{۲}{\delta}}{۲m}} \tag{۶۲.۲}$$

برهان اگر سمت راست معادله (۵۶.۲) را برابر δ قرار دهیم

$$۲e^{\left(-۲m\epsilon^۲\right)} = \delta \tag{۶۳.۲}$$

و سپس برای ϵ حل نماییم خواهیم داشت

$$\epsilon^۲ = \frac{\ln \frac{۲}{\delta}}{۲m} \tag{۶۴.۲}$$

اما ϵ کران پایین $\left|\hat{\mathbf{R}}(h) - \mathbf{R}(h)\right|$ در قضیه ۶.۲ است. لذا داریم

$$\left|\hat{\mathbf{R}}(h) - \mathbf{R}(h)\right|^۲ \geq \epsilon^۲ = \frac{\ln \frac{۲}{\delta}}{۲m} \tag{۶۵.۲}$$

$$\left|\hat{\mathbf{R}}(h) - \mathbf{R}(h)\right| \geq \sqrt{\frac{\ln \frac{۲}{\delta}}{۲m}} \tag{۶۶.۲}$$

$$\hat{\mathbf{R}}(h) - \mathbf{R}(h) \geq \sqrt{\frac{\ln \frac{۲}{\delta}}{۲m}} \tag{۶۷.۲}$$

و لذا خواهیم داشت

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}(h) + \sqrt{\frac{\ln \frac{۲}{\delta}}{۲m}} \tag{۶۸.۲}$$

■

آیا کران بدست آمده در نتیجه‌ی ۱.۲ را می‌توان برای فرضیه h که توسط الگوریتم یادگیری براساس داده‌های آموزشی S تولید می‌شود را تعمیم داد؟ پاسخ منفی است زیرا (الف) h یک فرضیه ثابت است و در اینجا توسط الگوریتم یادگیری مشخص می‌شود و (ب) h یک متغیر تصادفی است که به مجموعه آموزشی S وابسته است. بنابراین در نتیجه‌ی ۱.۲ دیگر رابطه $\mathbb{E}\left[\hat{\mathbf{R}}(h)\right] = \mathbf{R}(h)$ برقرار نیست زیرا $\mathbf{R}(h)$ یک متغیر تصادفی است و $\hat{\mathbf{R}}(h)$ یک عدد ثابت. در قضیه‌ای که در ادامه بیان می‌شود نتیجه‌ی ۱.۲ را برای این منظور تعمیم می‌دهیم.

قضیه ۷.۲ (کران خطا برای یادگیر بدون پیش فرض).
فرض کنید که H مجموعه‌ای متناهی از فرضیه‌ها باشد آنگاه برای هر $\delta > ۰$ با احتمال دست کم $(۱ - \delta)$، نابرابری زیر برای همه فرضیه‌های $h \in H$ برقرار است.

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}(h) + \sqrt{\frac{\log|H| + \log \frac{۲}{\delta}}{۲m}} \tag{۶۹.۲}$$

برهان فرض کنید که فرضیه‌های $\{h_۱, h_۲, \ldots, h_{|H|}\}$ اعضا مجموعه H باشند، آنگاه بر اساس کران اجتماعات و سپس اعمال نتیجه‌ی ۱.۲ خواهیم داشت

$$\begin{aligned}
\mathbb{P}\left[\exists h \in H : \left|\hat{\mathbf{R}}(h] - \mathbf{R}(h)\right| \geq \epsilon\right] &= \mathbb{P}\left[\left(\left|\hat{\mathbf{R}}(h_۱) - \mathbf{R}(h_۱)\right| \geq \epsilon\right) \vee \ldots \vee \left(\left|\hat{\mathbf{R}}(h_{|H|}) - \mathbf{R}(h_{|H|})\right| \geq \epsilon\right)\right] \\
&\leq \sum_{h_i \in H} \mathbb{P}\left[|\hat{\mathbf{R}}(h_i] - \mathbf{R}(h_i)| \geq \epsilon\right] \qquad (۷۰.۲) \\
&\leq ۲|H| \exp\left(-۲m\epsilon^۲\right) \leq \delta. \qquad (۷۱.۲)
\end{aligned}$$

و با حل معادله بالا براساس ϵ اثبات قضیه کامل می‌شود. ■

اگر نابرابری (۷۱.۲) را براساس m حل کنیم پیچیدگی نمونه‌ای یادگیری فضا بصورت زیر خواهدشد.

$$m \geq \frac{۱}{۲\epsilon^۲}\left[\log|H| + \log \frac{۲}{\delta}\right] \tag{۷۲.۲}$$

قضیه بالا نشان می‌دهد که هرچه m بزرگتر باشد خطای واقعی کمتر می‌شود و هرچه $|H|$ بزرگتر باشد خطای تجربی کمتر می‌شود اما خطای واقعی افزایش می‌یابد. بنابراین باید مصالحه‌ای بین فضای فرضیه و خطای واقعی صورت گیرد. این موضوع تیغ اوکام را نتیجه می‌دهد.

۳

بعد VC و پیچیدگی رادمیچرا

در بخش پیش نشان داده‌شد که اگر H فضای فرضیه متناهی باشد آنگاه ممکن است مساله قابل یادگیری باشد. این اثبات براساس متناهی بودن H بنا شده‌است. پیش از این مثال‌های مختلفی از فضاهای فرضیه همانند آستانه‌یک‌بعدی، مستطیل‌هایی با اضلاع موازی محورها، خط و بازه داشتیم که متناهی نبودند. اما در مدل‌های مختلف یادگیری، یادگرفتنی بودند . در نتیجه کران‌های بدست آمده در بخش پیش قابل تعمیم به فضای فرضیه‌های نامتناهی همانند فضای خط‌ها و فضای توابع چند جمله‌ای نیست. به همین دلیل اندازه فضای H مهم نیست بلکه قدرت توصیف یا غنی بودن این فضا مهم است.. برای اینکه بتوانیم از این فضای فرضیه‌ها استفاده نماییم می‌بایست جمله $\ln|H|$ در کران خطای واقعی یا پیچیدگی نمونه ای را با معیارهای دیگری که پیچیدگی یا غنا یا قدرت توصیف فضای فرضیه را نشان دهند جایگزین نماییم. برای اینکه شهود بهتری از پیچیدگی فرضیه‌ها را داشته باشیم فرض کنید که در فضای H، هر فرضیه با تابع k پارامتر حقیقی نمایش داده شود. برای نمونه در فضای فرضیه‌های مکعب مستطیل های n بعدی که اضلاع آن مواز محورها است $k = ۲n$ پارامتر و در یک ابر صفحه در فضای n بعدی $k = n + ۱$ پارامتر داریم. در این حالت ما می‌خواهیم این فرضیه‌ها را در حافظه محدود کامپیوتر نشان دهیم. اگر هر پارامتر را با b بیت نمایش دهیم نیاز به kb بیت برای نمایش تابع نیاز خواهیم داشت و در نتیجه فضای فرضیه H توان نمایش ۲^{kb} فرضیه متفاوت را خواهد داشت. برای سادگی فرض کنید که فرضیه پیدا شده با داده های آموزشی سازگار باشد در این‌صورت با توجه به اینکه فضای فرضیه دارای $|H| = ۲^{kb}$ فرضیه است و با جایگذاری این مقدار در کران پیچیدگی نمونه ای، پیچیدگی دارای کران پایین زیر خواهد بود.

$$m = O\left(\frac{۱}{\epsilon}\left[k + \ln\frac{۱}{\delta}\right]\right) \qquad (۱.۳)$$

در این حالت تعداد نمونه های مورد نیاز یک تابع خطی از تعداد پارامترها است. هر چند کران بالا یک رابطه خطی را برای پیچیدگی نمونه ای نشان می دهد اما گویای قدرت توصیف یا دسته بندی فضای فرضیه H نیست. در ادامه این بخش به بیان تعدادی معیار که توانایی

بیشتری در بیان پیچیدگی و قدرت توصیف فضای فرضیه دارند پرداخته می شود

۱.۳ تابع رشد

فرض کنید که H فضای فرضیه و $S = \{(x_۱, y_۱), (x_۲, y_۲), \ldots, (x_m, y_m)\}$ مجموعه آموزشی باشد. تابع رشد فضای فرضیه H برای مجموعه آموزشی S که با $\Pi_H(S)$ نشان داده می‌شود به صورت زیر تعریف می‌شود.

تعریف ۱.۳ **تابع رشد** تابع رشد فضای فرضیه H برای مجموعه آموزشی S برابر است با تعداد دسته بندی های ممکن S بوسیله H.

به عبارتی دیگر $\Pi_H(S)$ برابر است با تعداد بردارهای مجزای $(h(x_۱), h(x_۲), \ldots, h(x_m))$. لذا برای هر فضای فرضیه $h \in H$ و برای هر مجموعه آموزشی S می توان نشان داد که نتیجه زیر برقرار است.

نتیجه ۱.۳ .
هر فضای فرضیه H و برای مجموعه آموزشی S نامساوی $\Pi(S) \leq ۲^m$ برقرار است.

یک فاکتور مهم در نمایش قدرت توصیف فضای فرضیه‌ها بیشترین تعداد دسته بندی‌های ممکن است زیرا بسیاری از توابع در فضای H رفتارهای مشابه‌ای دارند. در نتیجه تابع رشد برابر است با

$$\Pi_H(m) = \max\{\Pi_H(s) \mid s \in \mathcal{X}^m\}. \tag{۲.۳}$$

به تابع $\Pi_H(S)$، تابع رشد می گویند. در ادامه چند مثال از توابع رشد برای فضاهای فرضیه مختلف بررسی می‌شود.

مثال ۱.۳ (توابع آستانه یک بعدی).
فرض کنید که H مجموعه‌ای از توابع آستانه یک بعدی باشد. اگر مجموعه آموزشی دارای یک نقطه باشد به دو حالت می‌توان این نمونه را برچسب زد. لذا $\Pi_H(S) = ۲$. اگر مجموعه آموزشی دارای دو نقطه باشد به سه حالت می‌توان این نمونه را برچسب زد. لذا $\Pi_H(S) = ۳$. و اگر مجموعه آموزشی دارای m نقطه باشد تابع رشد برابر است با $\Pi_H(S) = m + ۱$. همان گونه که تابع رشد نشان می‌دهد مقدار این تابع بسیار کوچکتر از ۲^m است. فرض کنید که نمونه‌های آموزشی به صورت $x_۱ < x_۲ < \ldots < x_m$ مرتب شده باشند. حال برای یک مقدار $\theta \in \mathbb{R}$، $h_\theta(x) = ۱$ است اگر $x \geq \theta$. در نتیجه برای هر $k \in [۱, m-۱]$ اگر $h_\theta(x_k) = ۱$ باشد نتیجه می‌گیریم که $h_\theta(x_{k+۱}) = ۱$ است. در نتیجه بردار، $(h(x_۱), h(x_۲), \ldots, h(x_m))$. دارای $m + ۱$ مقدار زیر خواهد بود.

$$(۰, ۰, \ldots, ۰), (۰, ۰, \ldots, ۱), \ldots, (۰, ۱, \ldots, ۱), (۱, ۱, \ldots, ۱)$$

در نتیجه مقدار تابع رشد برای این فضا برابر است با

$$\Pi_H(S) = m + ۱.$$

همان گونه که روشن است اندازه این فضا بی‌نهایت است یعنی بی نهایت مقدار برای θ وجود دارد در حالیکه بسیاری از این فرضیه ها برای مجموعه آموزشی همانند هم هستند و قابل تفکیک نیستند. بنابراین تابع رشد یک پارامتر بهتری نسبت به تعداد فرضیه ها است.

مثال ۲.۳ (فضای فرضیه های نیم خط).
این کلاس فرضیه بسیار شبیه توابع آستانه یک بعدی است با این تفاوت که نمونه های با مقدار بزرگتر از θ می‌توانند برچسب مثبت یا منفی داشته‌باشند. در حالیکه در توابع آستانه یک بعدی نمونه‌های با مقدار بزرگتر از θ تنها می‌توانستند برچسب مثبت داشته باشند. بنابراین تعداد حالت‌های مجزا برای دسته بندی دو برابر تعداد حالت های مجزا برای دسته بندی در توابع آستانه یک بعدی است. اگر تعداد حالت هایی که دو بار شمرده شده اند را کم نماییم خواهیم داشت.

$$\Pi_H(m) = ۲(m+۱) - ۲ = ۲m. \tag{۳.۳}$$

مثال ۳.۳ (فضای فرضیه های بازه).
فرض کنید که H مجموعه ای از توابع بازه باشد باشد. یا به عبارتی $H = \{[a,b] \mid a < b \in \mathbb{R}\}$. یک نمونه به صورت زیر دسته بندی می‌شود.

$$h(x) = \begin{cases} ۱ & if x \in [a,b] \\ ۰ & if x \notin [a,b] \end{cases} \tag{۴.۳}$$

اگر m نمونه آموزشی داشته باشیم $m+۱$ حالت برای قراردادن هرکدام از مرزها خواهیم داشت. در نتیجه $\Pi_H(m) = \binom{m+۱}{۲}+۱$. حالت $+۱$ زمانی رخ می دهد که بازه بین دو نمونه متوالی قرار بگیرند و در نتیجه خواهیم داشت

$$\Pi_H(m) = \binom{m+۱}{۲} + ۱ = O\left(m^۲\right) \tag{۵.۳}$$

که باز هم این مقدار بسیار کمتر از ۲^m است.

در بدترین حالت به یک فضای فرضیه H نیاز خواهیم داشت که $\Pi_H(S) = ۲^m$. از آنجایی که شرط $\Pi_H(S) < |H|$ برقرار است می‌توانیم با بکار بردن $\Pi_H(S)$ به جای H کران سفت تری را برای پیچیدگی نمونه‌ای پیدا کنیم و لذا قضیه زیر را می‌توان اثبات نمود.

قضیه ۱.۳ (کران خطا برای فرضیه های سازگار).
فرض کنید H خانواده ای از توابع باشد که مقدار $\{-۱,+۱\}$ را می پذیرند آنگاه برای همه $h \in H$ های سازگار، برای هر $\delta > ۰$، با احتمال دست کم $۱ - \delta$ رابطه زیر برقرار است.

$$\mathbf{R}(h) = O\left(\frac{\ln \Pi_H(۲m) + \ln \frac{۲}{\delta}}{m}\right). \tag{۶.۳}$$

برهان اثبات این قضیه را به عنوان تمرین انجام دهید. ■

قضیه ۲.۳ (کران خطا برای فرضیه های ناسازگار).
فرض کنید H خانواده ای از توابع باشد که مقدار $\{-۱,+۱\}$ را می‌پذیرند آنگاه برای همه $h \in H$ ها، برای هر $\delta > ۰$، با احتمال دست کم $۱ - \delta$ رابطه زیر برقرار است.

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}(h) + \sqrt{\frac{۲\ln \Pi_H(m)}{m}} + \sqrt{\frac{\ln \frac{۱}{\delta}}{۲m}}. \tag{۷.۳}$$

برهان اثبات این قضیه را به عنوان تمرین انجام دهید. ■

۲.۳ بعد VC

در بخش پیش تابع رشد بیان گردید و اشاره شد که تابع رشد معیار مناسب‌تری نسبت به اندازه فضای فرضیه بود و کران سفت تری را نسبت به آن نتیجه می‌دهد یعنی تعداد دسته بندهای مختلفی که یک کلاس فرضیه می‌تواند ایجاد نماید یک معیار خوب است. همچنین تعداد دسته بندی‌های ممکن بوسیله H برای هر مجموعه آموزشی با اندازه m و داده های دو به دو مجزا حداکثر ۲^m است. اما مثال‌های بیان شده نشان دادند که معمولا $\Pi_H(m) \leq ۲^m$ است و محاسبه تابع رشد برای بسیاری از فضاها سخت و پیچیده می باشد. حال اگر ما حالتی را فرض کنیم که H امکان ایجاد همه ۲^m دسته بندی ممکن را داشته باشد در اینصورت اندازه مجموعه آموزشی تغییر می‌کند. اما نشان داده شد که اگر اندازه تابع رشد از ۲^m کمتر باشد کلاس فرضیه مورد نظر نمی‌تواند برخی از برچسب گذاری های ممکن را پوشش دهد. بنابر این یک معیار خوب می‌تواند این باشد که یک کلاس فرضیه توانایی دسته بندی تمام برچسب گذاری های ممکن یک مجموعه آموزشی با چه اندازه را خواهد داشت. این نوع نگرش معیاری دیگر را نتیجه می‌دهد که نام آن بعد VC[1] است و در این بخش به معرفی آن می‌پردازیم. این معیار همانند تابع رشد نیز یک معیار پیچیدگی کاملا ترکیبیاتی است و محاسبه آن نیز از تابع رشد راحت‌تر است. برای تعریف بعد VC ابتدا باید مفاهیم دورستگی [2] و تفکیک [3] را مشخص کرد. مجموعه فرضیه‌های H را در نظر بگیرید، یک دورستگی مجموعه S یکی از راه‌های برچسب‌گذاری داده‌های S بر اساس یکی از فرضیه‌های H است. همچنین گوییم مجموعه فرضیه‌های H مجموعه S با $m \geq ۱$ نمونه را تفکیک کرده‌است، اگر H همه دورستگی‌های ممکن از S را بپوشاند، یعنی $\Pi_H(m) = ۲^m$. در ادامه این بخش نخست چند تعریف و مثال آورده می‌شود.

تعریف ۲.۳ **دورستگی** دورستگی مجموعه S با اندازه $m > ۱$ بوسیله فضای فرضیه H عبارت است از دسته بندی یکی از برچسب گذاری‌های ممکن S به دو دسته.

تعریف ۳.۳ **تفکیک داده** همه دورستگی‌های مجموعه S با اندازه $m > ۱$ بوسیله فضای فرضیه H ایجاد می شود اگر H همه برچسب گذاری‌های ممکن یا افرازهای ممکن S را امکان پذیر نماید. یعنی داشته باشیم $\Pi_H(m) = ۲^m$.

براساس دو تعریف دورستگی و تفکیک داده ها بعد VC فضای فرضیه H به صورت زیر تعریف می‌شود.

تعریف ۴.۳ **بعد VC** بعد VC فضای فرضیه H عبارت است از اندازه بزرگترین مجموعه‌ای که توسط H به طور کامل تفکیک شود و به عبارتی

$$VC(H) = \max\{m \mid \Pi_H(m) = ۲^m\} \qquad (۸.۳)$$

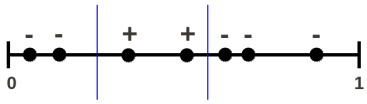
بنابراین اگر شرط $VC(H) = d$ برقرار باشد آنگاه مجموعه‌ای از اندازه d وجود دارد که همه دورستگی‌های آن توسط H شکسته می‌شود. اما به این معنی نیست که دورستگی‌های همه مجموعه‌های از اندازه d و یا کمتر توسط H شکسته می‌شوند. برای محاسبه بعد VC

[1] Vapnik-Chervonenkis Dimension
[2] dichotomy
[3] shattering

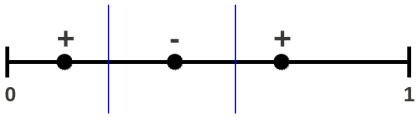
معمولا یک کران پایین برای آن اثبات می‌شود و برای این کار باید نشان داد که همه دوبخشی‌های مجموعه‌ای مانند S از اندازه d توسط H شکسته می‌شود. برای پیدا کردن یک کران بالا برای بعد VC باید نشان داد که هیچ مجموعه‌ای از اندازه $d+1$ توسط H شکسته نمی‌شود.

مثال ۴.۳ (بعد VC آستانه یک بعدی).
فضای فرضیه آستانه یک بعدی را در نظر بگیرید. این فضا تنها می‌تواند یک نمونه را تفکیک کند و لذا $VC(H) = 1$ خواهد بود.

مثال ۵.۳ (بازه‌های روی یک خط).
کران پایین بعد VC برای این مجموعه فرضیه ۲ است. برای مشاهده این نکته می‌توان همه دسته بندی های دو دسته ای دو نقطه را در نظر گرفت که در شکل ۱.۳ نشان داده شده‌است که توسط یک بازه از هم تفکیک شده‌اند. همچنین در شکل ۲.۳ با یک مثال نشان داده شده‌است که هیچ مجموعه سه تایی از نقاط را نمی‌توان توسط این مجموعه فرضیه از هم تفکیک کرد. بنابراین $VC(H) = 2$.



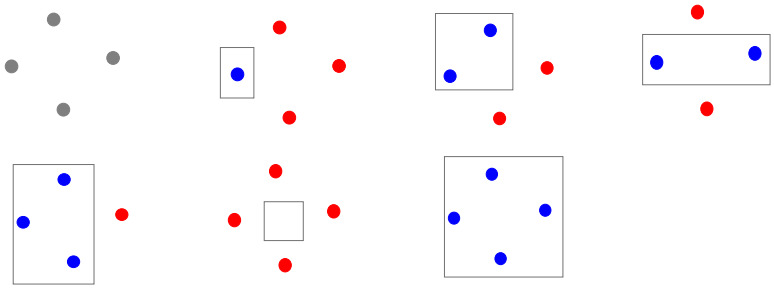
شکل ۱.۳ نمونه ای از نقاط که می توان با یک بازه آنها را از هم جدا کرد.



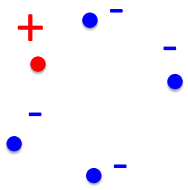
شکل ۲.۳ سه نقطه ای را که نمی توان با یک بازه آنها را از هم جدا کرد.

مثال ۶.۳ (مستطیل‌های با اضلاع موازی محورهای مختصات).
برای این مجموعه از فرضیه‌ها می‌توان نشان داد که کران پایین بعد VC دست کم ۴ است. برای این منظور ۴ نقطه روی یک لوزی درنظربگیریدکه اضلاع آن موازی محورها نیستند. همه حالت‌های دسته بندی دوتایی از این نقاط را می‌توان توسط یک مستطیل تشخیص داد. شکل ۳.۳ یک نمونه دیگر ای ۴ نقطه را نشان می‌دهد که می توانیم آنها را با یک مستطیل تفکیک کنیم. اما برای هر ۵ نقطه در صفحه اگر کوچکترین مستطیل دربردارنده همه نقاط را در نظر بگیرید، همواره یک نقطه درون این مستطیل قرار خواهد گرفت و با دادن برچسب $-$ به آن و دادن برچسب $+$ به مابقی نقاط مطابق شکل ۴.۳ هرگز نمی‌توان مستطیلی با اضلاع موازی محورها یافت که این کلاس‌ها را از هم تفکیک کند. بنابراین بعد VC برای این مجموعه فرضیه‌ها برابر با ۴ است.

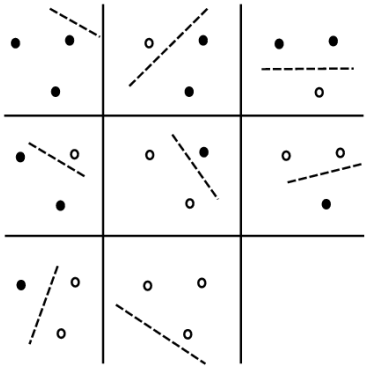
مثال ۷.۳ (بعد VC خطوط در فضای دو بعدی).
فضای فرضیه خطوط در فضای دو بعدی را در نظر بگیرید. مطابق شکل ۵.۳ دست کم یک سه نقطه‌ای وجود دارد که می‌توانیم ۳ نقطه در فضای دوبعدی را با یک خط تفکیک کنیم. بنابراین حداقل بعد VC برابر با ۳ است اما مطابق شکل نمی توانیم یک ۴ نقطه‌ای را درفضای دوبعدی بپیدا کنیم که با یک خط از هم تفکیک نماییم. بنابرای بعد VC خط در فضای دوبعدی کمتر از ۵ است و در نتیجه بعد VC خط در فضای دوبعدی مساوی ۴ است.



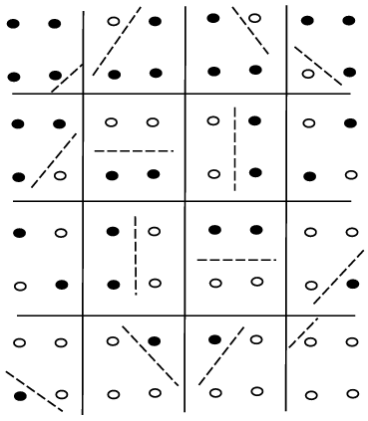
شکل ۳.۳ نمونه ای از ۴ نقطه که می توان با یک مستطیل از هم جدا کرد.



شکل ۴.۳ نمونه ای از ۵ نقطه که نمی توان با یک مستطیل از هم جدا کرد.



شکل ۵.۳ نمونه ای از سه نقطه که می توان با یک خط آنها را از هم جدا کرد.



شکل ۶.۳ چهار نقطه ای را که نمی توان با یک خط آنها را از هم جدا کرد.

قضیه ۳.۳ (کران بعد VC برای فضای فرضیه های متناهی).
اگر H یک فضای فرضیه متناهی باشد. آنگاه خواهیم داشت $VC(H) \leq \log |H|$.

برهان فرض کنید که $VC(H) = d$ باشد. بنابراین رابطه زیر برقرار است.

$$\Pi_H(d) = ۲^d. \tag{۹.۳}$$

اما برای هر مجموعه آموزشی با اندازه $m > ۱$ داریم $\Pi_H(m) \leq |H|$. و با توجه به رابطه بالا داریم $۲^d = \Pi_H(d) \leq |H|$. اگر از دو طرف رابطه بالا لگاریتم بگیریم قضیه اثبات می شود. ■

مثال ۸.۳ (بعد VC ترکیب عطفی حداکثر n متغیر).
برای فضای فرضیه H که از ترکیب عطفی حداکثر n ویژگی یا نقیض آن ساخته می شود رابطه زیر برقرار است.

$$n \leq VC(H) \leq n \log ۳. \tag{۱۰.۳}$$

درحالت کلی برای هر فضای برداری می‌توان نشان داد که لم زیر برقرار است.

لم ۱.۳ (بعد VC فضای برداری).
فرض کنید $\mathcal{G}$ مجموعه‌ای از توابع در فضای برداری $\mathbb{R}^d$ برای $d < \infty$ باشد. حال اگر مجموعه $\mathcal{F}$ را به صورت

$$\mathcal{F} = \{x \mapsto \text{sgn}\,(g(x)) \mid g \in \mathcal{G}\}$$

تعریف کنیم. اگر بعد $\mathcal{G}$ برابر با k باشد آنگاه بعد VC این فضا برابر است با

$$VC(\mathcal{F}) \leq k.$$

برهان اثبات این لم را به عنوان تمرین انجام دهید. ■

لم ۲.۳ (لم Sauer).
اگر H فرضیه‌ای با $VC(H) = d$، آنگاه برای هر $m \in \mathbb{N}$ رابطه زیر برقرار است:

$$\Pi_H(m) \leq \sum_{i=۰}^{d} \binom{m}{i} \tag{۱۱.۳}$$

برهان اثبات این لم به کمک استقرا روی $d + m$ انجام می‌گیرد. شما به عنوان تمرین آن را انجام دهید. ■

نتیجه ۲.۳ .
اگر H مجموعه‌ای از فرضیه‌ها باشد به طوری که $VC(H) = d$، آنگاه برای هر $m \geq d$ داریم

$$\Pi_H(m) \leq \left(\frac{em}{d}\right)^d = O(m^d) \tag{۱۲.۳}$$

برهان بااستفاده از لم ۲.۳ داریم

$$\begin{aligned}
\Pi_H(m) &\leq \sum_{i=0}^{d} \binom{m}{i} && (۱۳.۳)\\
&\leq \sum_{i=0}^{d} \binom{m}{i} \underbrace{\left(\frac{m}{d}\right)^{d-i}}_{>1} && (۱۴.۳)\\
&\leq \sum_{i=0}^{m} \binom{m}{i} \left(\frac{m}{d}\right)^{d-i} && (۱۵.۳)\\
&= \left(\frac{m}{d}\right)^d \sum_{i=0}^{m} \binom{m}{i} \left(\frac{d}{m}\right)^i \quad \text{براساس توزیع دو جمله ای} && (۱۶.۳)\\
&= \left(\frac{m}{d}\right)^d \left(1 + \frac{d}{m}\right)^m \quad (1-x) \leq e^{-x} \text{ براساس نامساوی} && (۱۷.۳)\\
&\leq \left(\frac{m}{d}\right)^d \left(e^{d/m}\right)^m && (۱۸.۳)\\
&= \left(\frac{m}{d}\right)^d e^d = \left(\frac{me}{d}\right)^d && (۱۹.۳)
\end{aligned}$$

■

نتیجه ۳.۳ .
اگر H مجموعه‌ای از توابع باشد که مقادیر آن از $\{-1, +1\}$ و $VC(H) = d$ باشد. آنگاه برای هر $\delta > 0$ با احتمال دست کم $1 - \delta$ برای هر $h \in H$ داریم:

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}(h) + \sqrt{\frac{2d \log \frac{em}{d}}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}} \tag{۲۰.۳}$$

برهان از کران تابع رشد داشتیم

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}(h) + \sqrt{\frac{2 \ln \Pi_H(m)}{m}} + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}}. \tag{۲۱.۳}$$

همچنین با استفاده از نتیجه لم ۲.۳ خواهیم داشت

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}(h) + \sqrt{\frac{2 \ln \Pi_H(m)}{m}} + \sqrt{\frac{\ln \frac{1}{\delta}}{2m}} \tag{۲۲.۳}$$

$$\leq \hat{\mathbf{R}}(h) + \sqrt{\frac{۲\ln\left(\frac{me}{d}\right)^d}{m}} + \sqrt{\frac{\ln\frac{۱}{\delta}}{۲m}} \tag{۲۳.۳}$$

$$\leq \hat{\mathbf{R}}(h) + \sqrt{\frac{۲d\ln\frac{me}{d}}{m}} + \sqrt{\frac{\ln\frac{۱}{\delta}}{۲m}}. \tag{۲۴.۳}$$

■

بنابراین شکل کلی کران به صورت زیر است:

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}(h) + O\left(\sqrt{\frac{\log(m/d)}{(m/d)}}\right) \tag{۲۵.۳}$$

نتیجه‌گیری ۲.۳ بسیار مهم است زیر این نتیجه‌گیری می‌گوید که همه کلاس‌های فرضیه ها در یکی از دو گروه زیر قرار می‌گیرند.

۱. اگر d نامتناهی باشد آنگاه $\Pi_H(m) = ۲^m$ است و در نتیجه کران زیر

$$\mathbf{R}(h) = O\left(\frac{\ln\Pi_H(۲m) + \ln\left(\frac{۲}{\delta}\right)}{m}\right). \tag{۲۶.۳}$$

بدون معنی می باشد.

۲. اگر d متناهی باشد آنگاه رابطه $\Pi_H(m) = O\left(m^d\right)$ برقرار است. در این حالت قضیه یاد شده نتایج جالبی دارد. از آنجایی که $\ln\Pi_H(m) = O(d\ln m)$ ، لذا کران خطا به صورت خطی به d و به صورت وارون به m وابسته است و با افزایش m به سمت صفر میل می کند.

از لم ۲.۳ به سادگی می توان قضیه زیر را نتیجه گرفت.

قضیه ۴.۳ ().
اگر $VC(H) = d$ آنگاه برای همه فرضیه های سازگار درون H، با احتمال دست کم $۱ - \delta$ نابرابری زیر برقرار است.

$$\mathbf{R}(h) = O\left(\frac{d\ln m + \ln\left(\frac{۲}{\delta}\right)}{m}\right) \tag{۲۷.۳}$$

و به طور مشابه می توان نتیجه گرفت که

$$m = O\left(\frac{۱}{\epsilon}\log\frac{۱}{\delta} + \frac{d}{\epsilon}\log\frac{۱}{\epsilon}\right). \tag{۲۸.۳}$$

در نتیجه پیچیدگی نمونه ای به صورت خطی با d رشد می نماید. برای نمونه می توانیم با افزودن تعداد ویژگی ها، بعد $VC(H)$ را افزایش دهیم. در نتیجه یادگیری هر بعد به تعداد ثابتی نمونه نیاز دارد. برای روشن شدن نتیجه بیان شده، چند مثال بیان می کنیم.

مثال ۹.۳ (تابع آستانه یک بعدی).
پیش از این نشان دادیم که برای فضای فرضیه آستانه یک بعدی $VC(H) = ۱$ است. بنابراین با جایگزین نمودن آن در نتیجه

قضیه قبل داریم

$$m = O\left(\frac{۱}{\epsilon}\log\frac{۱}{\delta} + \frac{۱}{\epsilon}\log\frac{۱}{\epsilon}\right). \tag{۲۹.۳}$$

اما پیش از این برای همین فضای فرضیه نشان دادیم که

$$m \geq \frac{۱}{\epsilon}\ln\frac{۲}{\delta}. \tag{۳۰.۳}$$

دو نتیجه بالا نشان می دهند که کرانی که بوسیله بعد VC پیدا می شود بد نیست.

مثال ۱۰.۳ (ابرصفحه در فضای n بعدی).
پیش از این نشان دادیم که بعد VC ابرصفحه n بعدی برابر $n+۱$ است. در نتیجه داریم

$$m = O\left(\frac{۱}{\epsilon}\log\frac{۱}{\delta} + \frac{n+۱}{\epsilon}\log\frac{۱}{\epsilon}\right). \tag{۳۱.۳}$$

این کران در واقع همان نتیجه گیری بدی را نشان می دهد که پیش از این در باره آن سخن به میان آمد. این رابطه نشان می دهد که پیچیدگی نمونه ای، یک تابع خطی از تعداد پارامترهای فرضیه یا همان تعداد ویژگی های نمونه است. حال مثال زیر را در نظر بگیرید.

مثال ۱۱.۳ (مستطیل های با اضلاع موازی محورها).
پیش از این نشان دادیم که برای فضای فرضیه مستطیل هایی که اضلاع آن موازی محورهای مختصات است $VC(H) = ۴$ است. بنابراین با جایگزین نمودن آن در نتیجه قضیه قبل داریم

$$m = O\left(\frac{۱}{\epsilon}\log\frac{۱}{\delta} + \frac{۴}{\epsilon}\log\frac{۱}{\epsilon}\right). \tag{۳۲.۳}$$

اما پیش از این برای همین فضای فرضیه نشان دادیم که

$$m \geq \frac{۴}{\epsilon}\ln\frac{۴}{\delta}. \tag{۳۳.۳}$$

حال مثال زیر را در نظر بگیرید که نتیجه بد گذشته را نقض می‌کند.

مثال ۱۲.۳ (فضای فرضیه).
فرض کنید که H فضای فرضیه $\text{sgn}(\sin(\theta x))$ باشد. می توان نشان داد که بعد $VC(H) = \infty$ است. اما این فرضیه تنها یک پارامتر دارد و نتیجه بد گذشته را نقض می‌کند.

کران‌های ارایه‌شده در این بخش را می‌توان به حالت نحقق‌ناپذیر نیز گسترش داد. در این حالت، این کران به صورت زیر خواهد شد (در روابط به‌جای $\frac{۱}{\epsilon}$ مقدار $\frac{۱}{\epsilon^۲}$ قرار می‌گیرد).

$$m = O\left(\frac{۱}{\epsilon^۲}\log\frac{۱}{\delta} + \frac{d}{\epsilon^۲}\log\frac{۱}{\epsilon}\right). \tag{۳۴.۳}$$

رابطه (۲۸.۳) کران بالای پیچیدگی نمونه‌ای را نشان می‌دهد که به صورت خطی به بعد VC وابسته است. آیا این وابستگی خطی سفت

است؟ برای پاسخ به این پرسش می‌بایست کران پایین پیچیدگی نمونه‌ای را پیدا نمود. می توان نشان داد که دست کم $\frac{d}{۲}$ نمونه لازم است تا برای هر $\epsilon, \delta < \frac{۱}{۸}$ با احتمال دست کم $۱ - \delta$ نابرابری $\mathbf{R}(h) \leq \epsilon$ برقرار است. این کران پایین را می توان به صورت $\Omega(d/\epsilon)$ نشان داد. در نتیجه کران (۲۸.۳) سفت است و علاوه بر وابستگی خطی پیچیدگی نمونه ای به d، به صورت خطی به $\frac{۱}{\epsilon}$ نیز وابسته است. این نتیجه در قضیه زیر نشان داده شده است

قضیه ۵.۳ ().
فرض کنید $VC(H) = d$ باشد. برای هر الگوریتم یادگیری A، یک فرضیه h و یک توزیع $\mathcal{D}$ وجود دارد که اگر الگوریتم یادگیری A، مجموعه آموزشی با $m \leq \frac{d}{۲}$ نمونه که از توزیع $\mathcal{D}$ به صورت مستقل نمونه برداری شده باشند و توسط تابع c برچسب گذاری شده باشند را دریافت و فرضیه h را تولید نماید آنکاه نابرابری زیر برقرار است.

$$\mathbb{P}\left[\mathbf{R}(h) > \frac{۱}{۸}\right] > \frac{۱}{۸}. \tag{۳۵.۳}$$

همچنین نشان داده شده است که یک فضای فرضیه H و یک توزیع $\mathcal{D}$ وجود دارد بطوریکه $\Omega\left(\frac{۱}{\epsilon}\left[d \ln \frac{۱}{\epsilon} + \ln \frac{۱}{\delta}\right]\right)$ نمونه نیاز است تا برای هرفرضیه سازگار $h \in H$ نابرابری $\mathbf{R}(h) \leq \epsilon$ برقرار باشد. همچنین نشان داده شده است که برای هر $k > ۱$، الگوریتمی وجود دارد که تنها

$$O\left(\frac{۱}{\epsilon} d \log^{(k)} \frac{۱}{\epsilon} + \frac{۱}{\epsilon} \log \frac{۱}{\delta}\right) \tag{۳۶.۳}$$

نمونه نیاز دارد تا با احتمال دست کم $۱ - \delta$ کران بالای خطای واقعی آن ϵ باشد که $\log^{(k)}(x) = \log(\log(\ldots \log(x)))$ است که log در رابطه k بار تکرار شده است. در رابطه بالا پیچیدگی نمونه ای به k وابسته است.

۳.۳ پیچیدگی رادامیچر

پیش از این مجموعه آموزشی $\mathsf{S} = \{(x_۱, y_۱), (x_۲, y_۲), \ldots, (x_m, y_m)\}$ و فرضیه $h : \mathcal{X} \mapsto \{۰, ۱\}$ را در نظر گرفتیم. در این حالت خطای آموزشی برابر است با

$$\hat{\mathbf{R}}(h) = \frac{۱}{m} \sum_{i=۱}^{m} \mathbb{I}\left[h(x_i) \neq y_i\right] \tag{۳۷.۳}$$

حال اگر مجموعه برچسب ها را به صورت $\mathcal{Y} = \{-۱, +۱\}$ در نظر گیریم فرضیه h به صورت فرضیه $h : \mathcal{X} \mapsto \{-۱, +۱\}$ خواهد بود. براین اساس خطای آموزشی را می توانیم به صورت زیر تعریف نماییم.

$$\hat{\mathbf{R}}(h) = \frac{۱}{m} \sum_{i=۱}^{m} \mathbb{I}\left[h(x_i) \neq y_i\right] \tag{۳۸.۳}$$

$$= \frac{۱}{m} \sum_{i=۱}^{m} \begin{cases} ۱ & \text{if } (h(x_i), y_i) = (+۱, -۱) \text{ or } (h(x_i), y_i) = (-۱, +۱) \\ ۰ & \text{if } (h(x_i), y_i) = (+۱, +۱) \text{ or } (h(x_i), y_i) = (-۱, -۱) \end{cases} \tag{۳۹.۳}$$

$$= \frac{۱}{m} \sum_{i=۱}^{m} \frac{۱ - y_i h(x_i)}{۲} \tag{۴۰.۳}$$

$$= \frac{۱}{۲} - \frac{۱}{۲m}\sum_{i=۱}^{m} y_i h(x_i) \tag{۴۱.۳}$$

عبارت $\frac{۱}{m}\sum_{i=۱}^{m} y_i h(x_i)$ را می توان همبستگی برچسب واقعی y_i با برچسب پیش بینی شده $h(x_i)$ دانست. بنابراین همبستگی به صورت زیر به خطای آموزشی وابسته است.

$$\mathrm{Corr}(h) = ۱ - ۲\hat{\mathbf{R}}(h) \tag{۴۲.۳}$$

برای اینکه فرضیه hای را پیدا کنیم که خطای آموزشی را کمینه نماید می توانیم فرضیه hای را پیدا کنیم که همبستگی را بیشینه نماید. یا به عبارتی

$$h = \underset{h\in H}{\mathrm{argmax}} \frac{۱}{m}\sum_{i=۱}^{m} y_i h(x_i). \tag{۴۳.۳}$$

حال فرض کنید که در یک آزمایش برچسب واقعی y_i با متغیر تصادفی رادامیچر σ_i که به صورت زیر تعریف می شود جایگزین گردد.

$$\sigma_i = \begin{cases} +۱ & \text{با احتمال } \frac{۱}{۲} \\ -۱ & \text{با احتمال } \frac{۱}{۲} \end{cases} \tag{۴۴.۳}$$

حال هدف این است که تابع زیر بیشینه گردد.

$$h = \underset{h\in H}{\mathrm{argmax}} \frac{۱}{m}\sum_{i=۱}^{m} \sigma_i h(x_i). \tag{۴۵.۳}$$

این تابع هدف به جای انتخاب فرضیه hای که همبستگی با برچسب را بیشینه نماید فرضیه hبه گونه ای پیدا می شود که همبستگی با نویز (σ_i) را بیشینه نماید. از آنجایی که فرضیه h به متغیر تصادفی σ_i وابسته است برای اندازه گیری همبستگی فضای H نسبت به σ امید ریاضی به صورت زیر محاسبه می شود.

$$\mathop{\mathbb{E}}_{\sigma}\left[\max_{h\in H} \frac{۱}{m}\sum_{i=۱}^{m} \sigma_i h(x_i)\right] \tag{۴۶.۳}$$

مقدار بالا به نوعی قدرت توصیف یا غنای فضای فرضیه H را نشان می دهد. براساس دو حالت حدی، کران عبارت بالا را بدست می آوریم.

۱. اگر فضای فرضیه H دارای یک فرضیه باشد در اینصورت همبستگی برابر صفر است. زیرا عملگر max دیگر معنایی نداشته و بنابراین تنها مجموع روی $\sigma_i h(x_i)$ می باشد. از آنجایی که متغیر تصادفی رادمیچر بطور میانگین در نیمی از حالت ها دارای مقدار $+۱$ و در نیمی از حالت ها دارای مقدار -۱ هستند بنابراین حاصل جمع به طور میانگین صفر است.

۲. اگر فضای فرضیه H دارای ۲^m فرضیه باشد در اینصورت همبستگی برابر یک است. زیرا دنباله $\sigma = (\sigma_۱, \sigma_۲ \ldots, \sigma_m)$ دارای ۲^m حالت است و در این حالت برای هر دنباله σ یک فرضیه وجود دارد لذا فضای فرضیه توانایی دسته بندی هر کدام از دنباله ها را دارد و در نتیجه مقدار عبارت بالا برابر یک می شود.

بنابراین امید ریاضی همبستگی در بازه $[۰, ۱]$ قرار دارد.

حال فرض کنید به جای $h : \mathcal{X} \mapsto \{-۱, +۱\}$ با یک مجموعه توابع حقیقی $\mathcal{F} = \{f \mid f : \mathcal{Z} \mapsto \mathbb{R}\}$ کار کنیم. جایگزین نمودن H با $\mathcal{F}$ یک خانواده از توابع $h : \mathcal{Z} \mapsto \mathbb{R}$ را نتیجه می دهد. حال برای مجموعه نمونه های $\mathsf{S} = (z_۱, z_۲, \ldots, z_m)$ با $z_i \in \mathcal{Z}$ پیچیدگی

رادمیچر به صورت زیر تعریف می شود.

$$\hat{\mathcal{R}}_S(\mathcal{F}) = \mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{F}} \frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z_i)\right] \tag{۴۷.۳}$$

که $\hat{\mathcal{R}}_S(\mathcal{F})$ پیچیدگی تجربی رادامیچر نامیده می شود و بطور میانگین همبستگی تابع $\mathcal{F}$ را با نویز نشان می دهد. ما می خواهیم همبستگی تابع $\mathcal{F}$ را نسبت به توزیع $\mathcal{D}$ روی $\mathcal{X}$ را پیدا نماییم. برای این کار امید ریاضی $\hat{\mathcal{R}}_S(\mathcal{F})$ روی تمام مجموعه داده های با طول m که براساس به توزیع $\mathcal{D}$ نمونه برداری شده اند را محاسبه می کنیم.

$$\mathcal{R}_m(h) = \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\left[\hat{\mathcal{R}}_S(\mathcal{F})\right] \tag{۴۸.۳}$$

که $\mathcal{R}_m(h)$ پیچیدگی رادامیچر نامیده می شود محاسبه پیچیدگی تجربی رادامیچر برای برخی از فضاها یک مساله ان ـ پی سخت است. یکی از برتری های رابطه بالا این است که توانایی مدل سازی مدل های دیگر را دارد و کران هایی تولید می شود که سفت تر و مناسب ترهستند.

قضیه ۶.۳ ().
فرض کنید که H خانواده ای از توابع از $\mathcal{X}$ به مجموعه $\{-1,+1\}$ باشند و توزیع $\mathcal{D}$ روی فضای ورودی X باشد. آنگاه برای هر $\delta > 0$ با احتمال دست کم $1-\delta$ روی یک مجموعه S با اندازه m که از توزیع $\mathcal{D}$ نمونه برداری شده است برای هر $h \in H$ نامساوی های زیر برقرار است.

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}(h) + \mathcal{R}_m(H) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \tag{۴۹.۳}$$

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}(h) + \hat{\mathcal{R}}_S(H) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}} \tag{۵۰.۳}$$

نامساوی دوم وابسته به داده های آموزشی است. محاسبه $\mathcal{R}_m(H)$ و $\hat{\mathcal{R}}_S(H)$ سخت است و در بخش های بعد تلاش می کنیم با روش های ترکیباتی این موضوع را پیوند بزنیم.

برهان

■

قضیه ۷.۳ ().
فرض کنید که $\mathcal{F}$ خانواده ای از توابع از $\mathcal{Z}$ به مجموعه $[0,1]$ را نگاشت می نمایند و مجموعه $S = (z_1, z_2, \ldots, z_m)$ که $z_i \sim \mathcal{D}$ باشد. اگر $\mathbb{E}[f] = \mathbb{E}_{z\sim D}[f(z)]$ و $\hat{\mathbb{E}}_S[f] = \frac{1}{m}\sum_{i=1}^{m} f(z_i)$ تعریف شوند. با احتمال دست کم $1-\delta$ برای تمامی $f \in \mathcal{F}$ نامساوی زیر برقرار است.

$$\mathbb{E}[f] \leq \hat{\mathbb{E}}_S[f] + 2\mathcal{R}_m(\mathcal{F}) + O\left(\sqrt{\frac{\ln\frac{1}{\delta}}{m}}\right) \tag{۵۱.۳}$$

$$\mathbb{E}[f] \leq \hat{\mathbb{E}}_S[f] + 2\hat{\mathcal{R}}_S(\mathcal{F}) + O\left(\sqrt{\frac{\ln\frac{2}{\delta}}{m}}\right) \tag{۵۲.۳}$$

برهان برای اثبات این قضیه، نخست کران $\mathbb{E}[f]-\hat{\mathbb{E}}_S[f]$ را برای تمامی $f\in\mathcal{F}$ و یا به طور معادل کران $\sup_{f\in\mathcal{F}}\left\{\mathbb{E}[f]-\hat{\mathbb{E}}_S[f]\right\}$ را پیدا می کنیم. $\phi(S)$ را به صورت زیر تعریف می کنیم

$$\phi(S)=\sup_{f\in\mathcal{F}}\left\{\mathbb{E}[f]-\hat{\underset{S}{\mathbb{E}}}[f]\right\}. \tag{۵۳.۳}$$

در رابطه بالا $\phi(S)$ یک متغیر تصادفی است و به مجموعه آموزشی S وابسته است. فرض کنید که $S=(z_1,z_2,\ldots,z_i,\ldots,z_m)$ و $S'=(z_1,z_2,\ldots,z'_i,\ldots,z_m)$ دو مجموعه آموزشی باشند که تنها در یک نمونه با هم اختلاف داشته باشند. حال در گام های زیر قضیه را اثبات می‌کنیم. در گام نخست نشان می دهیم که با احتمال دست کم $1-\delta$ نابرابری

$$\phi(S)\le\underset{S}{\mathbb{E}}[\phi(S)]+\sqrt{\frac{\ln\frac{2}{\delta}}{2m}}. \tag{۵۴.۳}$$

برقرار است. برای اثبات قضیه نابرابری McDiarmid را به‌کار می‌بریم. در نابرابری McDiarmid اگر برای همه i ها نابرابری زیر برقرار باشد

$$|f(z_1,z_2,\ldots,z_i,\ldots,z_m)-f(z_1,z_2,\ldots,z'_i,\ldots,z_m)|\le c_i \tag{۵۵.۳}$$

آنگاه نابرابری زیر برقرار خواهد بود.

$$\mathbb{P}\left[|f(S)-f(S')|\ge\epsilon\right]\le 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^m c_i^2}\right) \tag{۵۶.۳}$$

از تعریف $\phi(S)$ داریم.

$$\begin{aligned}\phi(S) &= \sup_{f\in\mathcal{F}}\left\{\mathbb{E}[f]-\hat{\underset{S}{\mathbb{E}}}[f]\right\} &(۵۷.۳)\\ &= \sup_{f\in\mathcal{F}}\left\{\mathbb{E}[f]-\frac{1}{m}\sum_{i=1}^m f(z_i)\right\}. &(۵۸.۳)\end{aligned}$$

از آنجایی که $f(z_i)\in[0,1]$ است بنابراین تغییر هریک از نمونه ها از z_i به z'_i در مجموعه آموزشی S مقدار $\frac{1}{m}\sum_{i=1}^m f(z_i)$ را حداکثر $\frac{1}{m}$ تغییر می دهد. لذا مقدار c_i در نابرابری McDiarmid برابر با $\frac{1}{m}$ است. اگر این مقدار را در نابرابری McDiarmid قرار دهیم خواهیم داشت.

$$\begin{aligned}\mathbb{P}\left[|\phi(S)-E_S[\phi(S)]|\ge\epsilon\right] &\le 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^m c_i^2}\right) &(۵۹.۳)\\ &= 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^m \left(\frac{1}{m}\right)^2}\right) &(۶۰.۳)\\ &= 2\exp\left(-2m\epsilon^2\right). &(۶۱.۳)\end{aligned}$$

اگر مقدار ϵ را برابر $\sqrt{\frac{\log 2\delta}{2m}}$ قرار دهیم با ساده سازی نابرابری بالا خواهیم داشت.

$$\phi(S)\le\underset{S}{\mathbb{E}}[\phi(S)]+\sqrt{\frac{\ln\frac{2}{\delta}}{2m}}. \tag{۶۲.۳}$$

در گام دوم ثابت می کنیم که نابرابری $\mathbb{E}_S[\phi(S)]\le 2\mathcal{R}_m(\mathcal{F})$ برقرار است. براساس تعریف $\phi(S)$ داریم

$$\underset{S}{\mathbb{E}}[\phi(S)] = \underset{S}{\mathbb{E}}\left[\sup_{f\in\mathcal{F}}\left(\mathbb{E}[f]-\hat{\underset{S}{\mathbb{E}}}[f]\right)\right] \tag{۶۳.۳}$$

$$= \mathop{\mathbb{E}}_{S}\left[\sup_{f\in\mathcal{F}}\mathop{\mathbb{E}}_{S'}\left[\mathop{\mathbb{E}}_{S'}[f]-\hat{\mathop{\mathbb{E}}}_{S}[f]\right]\right]. \tag{۶۴.۳}$$

از آنجایی که مجموعه S همانند مجموعه S' از توزیع $\mathcal{D}$ نمونه برداری شده اند لذا $\mathbb{E}[f]=\mathbb{E}_{S'}\left[\hat{\mathbb{E}}_{S'}[f]\right]$ است و در نتیجه می توانیم خط دوم معادله بالا را بنویسیم. با توجه به اینکه تابع sup یک تابع محدب است لذا می توانیم از نابرابری Jensen استفاده و معادله بالا را به صورت زیر بنویسیم.

$$\mathop{\mathbb{E}}_{S}[\phi(S)] \leq \mathop{\mathbb{E}}_{S,S'}\left[\sup_{f\in\mathcal{F}}\left(\hat{\mathop{\mathbb{E}}}_{S'}[f]-\hat{\mathop{\mathbb{E}}}_{S}[f]\right)\right] \tag{۶۵.۳}$$

$$= \mathop{\mathbb{E}}_{S,S'}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\left(f(z'_i)-f(z_i)\right)\right]. \tag{۶۶.۳}$$

اگر متغیر رادمیچر را به نابرابری بالا اضافه کنیم امید ریاضی آن تغییری نمی کند. زیرا در نابرابری بالا اگر مقدار $\sigma_i=1$ باشد نابرابری بالا تغییری نمی کند و اگر $\sigma_i=-1$ باشد جای z_i و z'_i را عوض می کنیم چون امید ریاضی را روی تمام S و S' ها محاسبه می شود لذا تغییری در امید ریاضی نخواهیم داشت. بنابراین با افزودن متغیر رادمیچر به نابرابری بالا، خواهیم داشت.

$$\mathop{\mathbb{E}}_{S}[\phi(S)] = \mathop{\mathbb{E}}_{\sigma,S,S'}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i\left(f(z'_i)-f(z_i)\right)\right]. \tag{۶۷.۳}$$

اگر از ویژگی $\sup(a+b)\leq\sup(a)+\sup(b)$ استفاده کنیم خواهیم داشت.

$$\mathop{\mathbb{E}}_{S}[\phi(S)] \leq \mathop{\mathbb{E}}_{\sigma,S'}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z'_i)\right]+\mathop{\mathbb{E}}_{\sigma,S}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}-\sigma_i f(z_i)\right]. \tag{۶۸.۳}$$

با توجه به اینکه متغیرهای رادمیچر σ_i و $-\sigma_i$ دارای توزیع یکسانی هستند لذا خواهیم داشت.

$$\mathop{\mathbb{E}}_{S}[\phi(S)] \leq 2\mathop{\mathbb{E}}_{\sigma,S}\left[\sup_{f\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i f(z_i)\right] \tag{۶۹.۳}$$

$$= 2\mathcal{R}_m(\mathcal{F}). \tag{۷۰.۳}$$

با جایگذاری نابرابری (۷۰.۳) در نابرابری (۵۴.۳) داریم

$$\phi(S) \leq 2\mathcal{R}_m(\mathcal{F})+\sqrt{\frac{\ln\frac{2}{\delta}}{2m}}. \tag{۷۱.۳}$$

که با جایگذاری، قضیه اثبات می گردد. ■

نابرابری (۵۲.۳) نیز به صورت مشابه اثبات می گردد. در ادامه این بخش، نخست به بررسی کاربردی از پیچیدگی رادمیچر می‌پردازیم.

لم ۳.۳ (کران رادمیچر برای خطای ۰ـ۱).
فرض کنید که فرضیه $h:\mathcal{X}\mapsto\{-1,+1\}$ یک دسته بند دودویی و H فضای فرضیه و همچنین $f_h(x,y)=\mathbb{I}[h(x)\neq y]$ تابع خطای صفر_یک و $\mathcal{F}_H=\{f_h\mid h\in H\}$ باشد. آنگاه برای هر مجموعه $S=\{(x_1,y_1),\ldots,(x_m,y_m)\}$ فرض کنید $S_X=(x_1,\ldots,x_m)$ تصویر S روی X باشد. آنگاه رابطه زیر بین پیچیدگی تجربی رادمیچر H و $\mathcal{F}$ برقرار است.

$$\hat{\mathcal{R}}_S(\mathcal{F}_H)=\frac{1}{2}\hat{\mathcal{R}}_{S_X}(H) \tag{۷۲.۳}$$

برهان برای اثبات این لم $Z = X \times \{-۱, +۱\}$ را در نظر بگیرید. براساس تعریف فضای $\mathcal{F}_H$، هر تابع $f_h \in \mathcal{F}_H$ متناظر با برخی از $h \in H$ است. از تعاریف مربوط به خطا داریم

$$\mathbf{R}(h) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} [\mathbb{I}[h(x) \neq y]] = \mathbb{E}[f_h] \tag{۷۳.۳}$$

$$\hat{\mathbf{R}}(h) = \frac{۱}{m}\sum_{i=۱}^{m} \mathbb{I}[h(x) \neq y] = \hat{\mathop{\mathbb{E}}_{S}}[f_h] \tag{۷۴.۳}$$

لذا مطابق قضیه ۷.۳، کران $\mathbf{R}(h) - \hat{\mathbf{R}}(h)$ را پیدا می کنیم.

$$\hat{\mathcal{R}}_S(F_H) = \mathop{\mathbb{E}}_{\sigma}\left[\sup_{f_h\in\mathcal{F}_H} \frac{۱}{m}\sum_{i=۱}^{m} \sigma_i f_h(x_i, y_i)\right] \tag{۷۵.۳}$$

$$= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in H} \frac{۱}{m}\sum_{i=۱}^{m} \sigma_i \left(\frac{۱ - y_i h(x_i)}{۲}\right)\right] \tag{۷۶.۳}$$

$$= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in H} \frac{۱}{۲m}\sum_{i=۱}^{m} \sigma_i + \sup_{h\in H} \frac{۱}{۲m}\sum_{i=۱}^{m} (-y_i\sigma_i) h(x_i)\right] \tag{۷۷.۳}$$

$$= \frac{۱}{۲m}\sum_{i=۱}^{m} \mathop{\mathbb{E}}_{\sigma}[\sigma_i] + \frac{۱}{۲}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in H} \frac{۱}{m}\sum_{i=۱}^{m} \sigma_i h(x_i)\right] \tag{۷۸.۳}$$

با توجه به ویژگی های متغیرهای تصادفی رادمیچر، جمله اول برابر با صفر و همچنین توزیع $-y_i\sigma_i$ همانند توزیع σ_i است لذا معادله بالا به صورت زیر خواهد شد.

$$\hat{\mathcal{R}}_S(F_H) = \frac{۱}{۲}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in H} \frac{۱}{m}\sum_{i=۱}^{m} \sigma_i h(x_i)\right] \tag{۷۹.۳}$$

$$= \frac{۱}{۲}\hat{\mathcal{R}}_{S_X}(H). \tag{۸۰.۳}$$

که بدین ترتیب اثبات لم کامل می گردد.

■

اگر نتیجه لم ۳.۳ را در نتایج قضیه ۷.۳ جایگزین نمایم کران خطای زیر را بدست خواهیم آورد.

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}(h) + \hat{\mathcal{R}}_{S_X}(H) + O\left(\sqrt{\frac{\ln\frac{۱}{\delta}}{m}}\right). \tag{۸۱.۳}$$

در ادامه رابطه بین پیچیدگی رادمیچر و معیارهای دیگری که پیش از این برای توصیف غنای فضای فرضیه بررسی نمودیم را بدست می آوریم.

قضیه ۸.۳ ().
برای هر فضای فرضیه $|H| < \infty$ نابرابری زیر برقرار است.

$$\hat{\mathcal{R}}_S(H) \leq \sqrt{\frac{۲\ln|H|}{m}}. \tag{۸۲.۳}$$

پیش از اثبات این قضیه، نخست لم زیر که به لم ماسارت معرف است را ارایه می کنیم.

لم ۴.۳ (لم ماسارت).
فرض کنید $A \subset \mathbb{R}^m$ مجموعه متناهی از بردارها باشد به گونه ای که برای هر بردار $\mathbf{a} \in A$، نابرابری $\|\mathbf{a}\| \leq ۱$ برقرار باشد. آنگاه نابرابری زیر برقرار است.

$$\mathop{\mathbb{E}}_{\sigma}\left[\max_{a\in A}\sum_{i=۱}^{m}\sigma_i a_i\right] \leq \sqrt{۲\ln|A|} \tag{۸۳.۳}$$

که σ_i متغیرهای تصادفی رادمیچر و $a_۱, \ldots, a_m$ مولفه های بردار $\mathbf{a}$ هستند.

اثبات قضیه ۸.۳ اگرفضای A را به صورت زیر تعریف کنیم.

$$A = \left\{\frac{۱}{\sqrt{m}}(h(x_۱), h(x_۲), \ldots, h(x_m))\right\}.$$

آنگاه A مجموعه ای از بردارها در فضای $\mathbb{R}^m$ است که برای همه $\mathbf{a} \in A$ داریم $\|a\| = ۱$. لذا از تعریف پیچیدگی تجربی رادمیچر داریم

$$\begin{aligned}
\hat{\mathcal{R}}_S(H) &= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in H}\frac{۱}{m}\sum_{i=۱}^{m}\sigma_i h(x_i)\right] && (۸۴.۳)\\
&= \mathop{\mathbb{E}}_{\sigma}\left[\sup_{\mathbf{a}\in A}\frac{\sqrt{m}}{m}\sum_{i=۱}^{m}\sigma_i a_i\right] && (۸۵.۳)\\
&= \frac{۱}{\sqrt{m}}\mathop{\mathbb{E}}_{\sigma}\left[\max_{\mathbf{a}\in A}\sum_{i=۱}^{m}\sigma_i a_i\right] && (۸۶.۳)\\
&\leq \frac{۱}{\sqrt{m}}\sqrt{۲\ln|A|} && (۸۷.۳)\\
&= \sqrt{\frac{۲\ln|A|}{m}}. && (۸۸.۳)
\end{aligned}$$

نابرابری بالا براساس لم ماسارت بدست آمده است. با توجه به اینکه مجموعه A تمام دسته بندی‌های ممکن را برای مجموعه S را ارایه می نماید لذا $A \subseteq H$ و در نتیجه $|A| \leq |H|$ را داریم و بنابراین قضیه اثبات می گردد. ■

مشکل کران بدست آمده در قضیه بالا این است که قید $|H| < \infty$ وجود دارد و برای بسیاری از فضاهای فرضیه کارایی ندارد. از آنجایی که $\hat{\mathcal{R}}_S(H)$ علاوه بر H تنها به S وابسته است می خواهیم مشخص نماییم که چگونه رفتار همه فرضیه های این فضا به مجموعه آموزشی وابسته هستند. در قضایای زیر این قید حذف شده و کران های سفت تری بدست می آوریم.

قضیه ۹.۳ ().
برای هر فضای فرضیه H داریم

$$\hat{\mathcal{R}}_S(H) \leq \sqrt{\frac{۲\ln \Pi_H(S)}{m}} \leq \sqrt{\frac{۲\ln \Pi_H(m)}{m}}. \tag{۸۹.۳}$$

برهان در اثبات این قضیه می خواهیم رفتار فرضیه ها به مجموعه آموزشی را مشخص نماییم. برای هر مجموعه آموزشی، فضای فرضیه H' را زیر مجموعه ای H تعریف نموده که برای هر رفتار S یک نماینده از H در نظر می گیرد. دقت شود که H' زیر مجموعه ای از H است و دارای ویژگی زیر است.

$$|H'| = \Pi_H(S) \leq \Pi_H(m) \leq ۲^m < \infty. \tag{۹۰.۳}$$

حال براساس تعریف پیچیدگی تجربی رادمیچر داریم

$$\hat{\mathcal{R}}_S(H) = \mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in H}\frac{۱}{m}\sum_{i=۱}^{m}\sigma_i h(x_i)\right] \tag{۹۱.۳}$$

از آنجایی که برای هر $h \in H$ که مقدار $\hat{\mathcal{R}}_S(H)$ را بیشینه نماید یک $h' \in H'$ وجود دارد که همان مقدار را نتیجه می دهد لذا داریم

$$\hat{\mathcal{R}}_S(H) = \mathop{\mathbb{E}}_{\sigma}\left[\sup_{h'\in H'}\frac{۱}{m}\sum_{i=۱}^{m}\sigma_i h'(x_i)\right] \tag{۹۲.۳}$$

$$= \hat{\mathcal{R}}_S(H'). \tag{۹۳.۳}$$

رابطه بالا نشان می دهد که sup روی H بیشتر از sup روی H' نخواهد بود و همچنین sup روی H' بیشتر از sup روی H نخواهد بود و بنابراین هر دو sup با هم برابر خواهند بود لذا از قضیه ۸.۳ داریم

$$\hat{\mathcal{R}}_S(H) = \hat{\mathcal{R}}_S(H') \tag{۹۴.۳}$$

$$\leq \sqrt{\frac{۲\ln|H'|}{m}} \tag{۹۵.۳}$$

$$= \sqrt{\frac{۲\ln\Pi_H(S)}{m}} \tag{۹۶.۳}$$

و به این ترتیب اثبات قضیه کامل می گردد. ■

حال رابطه پیچیدگی تجربی رادمیچر را با بعد VC بررسی می کنیم.

قضیه ۱۰.۳ ().
اگر $VC(H) = d$ باشد آنگاه برای هر $m \geq d \geq ۱$ داریم

$$\hat{\mathcal{R}}_S(H) \leq \sqrt{\frac{۲d\ln\left(\frac{em}{d}\right)}{m}} \tag{۹۷.۳}$$

برهان از لم ۲.۳ داشتیم

$$\Pi_H(m) \leq \left(\frac{em}{d}\right)^d. \tag{۹۸.۳}$$

با جایگزینی نابرابری بالا در نتیجه قضیه ۹.۳ داریم

$$\hat{\mathcal{R}}_S(H) \leq \sqrt{\frac{۲\ln\Pi_H(m)}{m}} \tag{۹۹.۳}$$

$$\leq \sqrt{\frac{۲\ln\left(\frac{em}{d}\right)^d}{m}} \tag{۱۰۰.۳}$$

$$= \sqrt{\frac{۲d\ln\left(\frac{em}{d}\right)}{m}} \tag{۱۰۱.۳}$$

$$= \sqrt{\frac{۲\ln\left(\frac{em}{d}\right)}{\left(\frac{m}{d}\right)}}. \tag{۱۰۲.۳}$$

و بدین ترتیب اثبات قضیه کامل می شود. ■

۴.۳ یادگیر های عمومی

پیش از این دیدیم که داده های آموزشی می توانند یادگیر را بد راهنمایی کنند و منجر به بیش برازش شود. یک راه برای حل این مشکل محدود کردن فضای جستجو به یک فضای فرضیه H است. این فضای فرضیه را می توان به صورت دانش پیشین دانست که در چنین فضایی یک فرضیه با خطای کم موجود است. آیا وجود چنین دانش پیشینی برای موفقیت یادگیر لازم است؟ یا به عبارتی دیگر می توان بدون دانش پیشین، یک یادگیر موفق داشت؟ به این یادگیر که از هیچ گونه دانش پیشینی استفاده نمی کند یادگیر عمومی می گویند.

در واقع پرسش را می توان این گونه مطرح نمود آیا برای یک مجموعه آموزشی با اندازه m، یک الگوریتم یادگیری A وجود دارد که برای هر توزیع $\mathcal{D}$، اگر الگوریتم A مجموعه آموزشی با m نمونه که از روی توزیع $\mathcal{D}$ به صورت مستقل نمونه برداری شده اند دریافت و با احتمال بالا فرضیه hی را تولید نماید که خطای کمی را تولید کند؟

قضیه No free lunch بیان می دارد که چنین یادگیر عمومی وجود ندارد. یا به بیان دقیق تر برای هر مساله دسته بندی، برای هر یادگیر یک توزیع احتمال وجود دارد که یادگیر روی آن توزیع توانایی یادگیری را نداشته و شکست می خورد. واژه شکست به این معنا است که یادگیر با دریافت مجموعه آموزشی با ویژگی های یاد شده، فرضیه ای را تولید می نماید که خطای زیادی (برای نمونه خطای ۲/۰) دارد در حالیکه برای این توزیع یادگیر دیگری وجود دارد که فرضیه ای با خطای کم را تولید می نماید. بنابراین در یک مساله یادگیری که نمونه ها با توزیع $\mathcal{D}$ نمونه برداری می شوند به دانش پیشین نیاز خواهیم داشت. این دانش پیشین می تواند گونه های مختلف زیر را داشته باشند.

- فرض کنیم که توزیع $\mathcal{D}$ یک شکل پارامتری داشته که پارامترهای آن نامشخص است. بنابراین می توانیم نخست پارامترهای توزیع را بدست آورده و سپس دسته بند را طراحی نماییم. به این گونه مدل ها، مدل های مولد می گویند.
- فرض کنیم یک فرضیه $h \in H$ وجود دارد به گونه ای که $\mathbf{R}(h) = ۰$. یک قید نرم تر روی دانش پیشین این است که $\min_{h\in H} \mathbf{R}(h)$ کوچک باشد.

قضیه ۱۱.۳ (No free lunch).
فرض کنید که A یک الگوریتم یادگیری برای دسته بندی دودویی با تابع خطای ۰/۱ در دامنه $\mathcal{X}$ باشد. همچنین فرض کنید که m عددی کوچک تر از $\frac{|\mathcal{X}|}{۲}$ باشد. آنگاه یک توزیع $\mathcal{D}$ روی $\mathcal{X} \times \{۰,۱\}$ وجود دارد به گونه ای که

۱. یک فرضیه $h : \mathcal{X} \mapsto \{۰,۱\}$ وجود دارد که $\mathbf{R}(h) = ۰$
۲. با احتمال دست کم $\frac{۱}{۷}$ روی مجموعه آموزشی S با اندازه m که از روی توزیع $\mathcal{D}$ نمونه برداری شده است داریم $\mathbf{R}(A(S)) \geq \frac{۱}{۸}$.

شیوه اثبات این قضیه به این صورت است که یک توزیع $\mathcal{D}$ را به گونه ای می سازیم که خطای فرضیه تولید شده زیاد باشد. نکته ای

که در این قضیه وجود دارد این است که هر الگوریتم که تنها نیمی از داده های فضا را ببیند نمی تواند در باره شیوه برچسب زنی داده ها داوری نماید. رابطه قضیه NFL با دانش پیشین چیست؟ برای بررسی این پرسش، فرض کنید $H = \{h \mid h : \mathcal{X} \mapsto \{۰, ۱\}\}$ مجموعه تمام توابع و A یک الگوریتم کمینه سازی خطای تجربی باشد. در واقع این نشان می دهد که هیچ گونه دانش پیشین نداریم. براین اساس نتیجه زیر حاصل می شود.

نتیجه ۴.۳ .
فرض کنید $\mathcal{X}$ فضای نمونه نامتناهی و $H = \{h \mid \mathcal{X} \mapsto \{۰, ۱\}\}$ مجموعه تمامی فرضیه های ممکن باشد آنگاه H قابلیت یادگیری احتمالا تقریبا درست را ندارد.

حال این پرسش مطرح می شود که چگونه می توان جلوی شکست الگوریتم یادگیری را گرفت؟ یک راه برای جلوگیری شکست الگوریتم یادگیری، محدود کردن فضای فرضیه H است. اما چگونه می توان فضای فرضیه خوب انتخاب نمود؟ پیش از این دیدم که انتخاب فضای فرضیه پیچیده راه حل مناسب نیست زیرا بیش برازش روی می دهد. از سویی دیگر، انتخاب فضای فرضیه ساده هم راه حل مناسب نیست چون خطای آموزشی زیادی دارد. بنابرای یک مصالحه وجود دارد. برای پاسخ به پرسش یاد شده، خطای الگوریتم کمینه سازی خطا را به دو بخش تجزیه می کنیم. برای این تجزیه خطا، فرض کنید الگوریتم کمینه سازی خطای تجربی با دریافت مجموعه آموزشی S فرضیه H را تولید می کند. بنابراین خطای واقعی h را می توان به صورت زیر نوشت.

$$\mathbf{R}(h) = \mathbf{R}_{app}(h) + \mathbf{R}_{est}(h) \qquad (۱۰۳.۳)$$

که $\mathbf{R}_{app}(h) = \min_{h \in H} \mathbf{R}(h)$ خطای تقریب و $\mathbf{R}_{est}(h) = \mathbf{R}(h) - \mathbf{R}_{app}(h)$ خطای برآورد نامیده می‌شود. خطای تقریب که عبارت است از کمترین خطا در فضای فرضیه H و براساس فضای فرضیه انتخاب شده مشخص می‌گردد و به مجموعه آموزشی وابسته نیست. خطای تقریب نشان می‌دهد چه مقدار از خطا براساس فرض‌های صورت گرفته برای محدود کردن به فضای فرضیه انتخاب شده ایجاد می‌شود. خطای برآورد عبارت است از اختلاف بین خطای تولید شده و خطای بهترین فرضیه در فضای فرضیه انتخاب شده. این خطا هم به مجموعه آموزشی و اندازه آن وابسته است و هم به پیچیدگی فضای فرضیه انتخاب شده. پیش از این نشان داده شد که این خطا برای فضای فرضیه متناهی به صورت لگاریتمی با $|H|$ افزایش و با افزایش m کاهش می‌یابد.

بهترین فضای فرضیه، فضایی است که تنها یک دسته بند دارد و آن دسته بند بهینه بیز می باشد. این دسته بند به توزیع $\mathcal{D}$ وابسته است که آن را نمی دانیم. یکی از اهداف نظریه یادگیری این است که چگونه بتوان یک فضای فرضیه غنی و در عین حال خطای برآورد را در حد معقولی داشت.

۴

یادگیری نایکنواخت

نماد یادگیری احتمالا تقریبا درست که پیش از این در باره آن سخن گفتیم اجازه می‌دهد که اندازه مجموعه آموزشی به پارامترهای دقت و اطمینان وابسته باشد اما این اندازه مجموعه آموزشی برای همه فرضیه‌ها و همه توزیع‌های داده برقرار است. کلاس‌های فرضیه‌هایی که در این مدل یادگرفتنی هستند محدود است و این کلاس‌ها می‌بایست دارای بعد VC متناهی باشند. در این بخش، یک مدل ضعیف تر و سست شده قابلیت یادگیری را مطرح می‌کنیم و ویژگی های این مدل یادگیری که یادگیری نایکنواخت نام دارد را بررسی می‌کنیم.

در ادامه، نخست مدل یادگیری نایکنواخت بررسی می‌شود که در این مدل، اندازه مجموعه آموزشی علاوه بر پارامترهای اطمینان و دقت به فرضیه نیز وابسته است. سپس ویژگی های این مدل را بررسی و نشان می‌دهیم که مدل نسخه سست شده مدل یادگیری احتمالا تقریبا درست بدون پیش فرض است و در ادامه شرط لازم برای اینکه فضای فرضیه به صورت نایکنواخت یادگرفتنی باشد را بیان می‌کنیم و یک الگوریتم یادگیری با نام کمینه سازی خطای ساختاری بیان می‌شود که توانایی یادگیری نایکنواخت را دارا است. همچنین روش یادگیری کمینه طول توصیف برای فضای فرضیه شمارا بیان می‌شود. در پایان مدل یادگیری سازگاری را که مدل ضعیف تر یادگیری است ارایه و برتری‌ها و کاستی‌های مدل‌های مختلف یادگیری که در فصل‌های پیشین ارایه شد را بررسی می‌کنیم.

۱.۴ قابلیت یادگیری نایکنواخت

پیش از این در باره همگرایی یکنواخت سخن گفتیم. واژه یکنواخت به این معنا به کار برده شد که با یک مجموعه آموزشی با اندازه ثابت و برای همه فرضیه‌های درون مجموعه فرضیه‌ها و برای همه توزیع‌ها، می‌توانستیم کران خطا یا پیچیدگی نمونه‌ای را تعیین نماییم. به بیانی دیگر همگرایی یکنواخت را می توان به صورت رسمی به صورت زیر بیان نمود.

تعریف ۱.۴ **همگرایی یکنواخت** کلاس فرضیه H ویژگی همگرایی یکنواخت را دارد اگر یک تابع $m_H^{UC} : (۰,۱)^۲ \mapsto \mathbb{N}$ و الگوریتم یادگیری A وجود داشته‌باشد به گونه‌ای که برای هر $\epsilon, \delta \in (۰,۱)$ و برای هر توزیع $\mathcal{D}$ روی $\mathcal{X} \times \mathcal{Y}$ اگر مجموعه آموزشی با اندازه $m > m_H^{UC}(\epsilon, \delta)$ به صورت مستقل از توزیع $\mathcal{D}$ از روی فضای نمونه، نمونه برداری شده باشند. آنگاه با احتمال دست کم $۱ - \delta$ نابرابری زیر برای همه فرضیه‌های درون H برقرار است.

$$\mathbf{R}(A(S)) \leq \mathbf{R}(h) + \epsilon. \tag{۱.۴}$$

فرض کنید که می‌خواهیم یک دسته‌بند طراحی نماییم که با احتمال بالا دارای خطای کمی باشد. اگر یک دسته‌بند خطی را در نظر گیریم ممکن است با احتمال بالا خطای آن کم نباشد. حال اگر دسته‌بند یک چندجمله‌ای از درجه ۲ باشد ممکن است خطای مناسب را نداشته‌باشد. اگر فضای فرضیه‌ها شامل همه چندجمله‌ای‌ها باشد به احتمال بسیار زیاد دسته‌بندی پیدا می‌کنیم که خطای آن کم است و لذا مناسب خواهد بود. اما بعد VC این فضا برابر ∞ است و بنابراین کران‌هایی که پیش از این بدست آوردیم دیگر ارزشی ندارد.

همان گونه که پیش از این بیان شد کلاس فرضیه‌هایی که قابلیت یادگیری با ویژگی همگرایی یکنواخت را داشته باشند بسیار محدود است و این کلاس فرضیه‌ها باید دارای بعد VC متناهی باشند. در ادامه شرایط بیان شده بالا را تعدیل می کنیم و اجازه می دهیم که اندازه مجموعه آموزشی علاوه بر ϵ و δ به h نیز وابسته باشد. به بیانی دیگر، اندازه مجموعه آموزشی به صورت یکنواخت برای همه فرضیه‌ها یکسان نیست. به این ویژگی، ویژگی همگرایی نایکنواخت یا یادگیری نایکنواخت می‌گوییم و به صورت زیر تعریف می‌شود.

تعریف ۲.۴ **یادگیری نایکنواخت** کلاس فرضیه H قابلیت یادگیری نایکنواخت را دارد اگر یک تابع $m_H^{NUL} : (۰,۱)^۲ \times H \mapsto \mathbb{N}$ و یک الگوریتم یادگیری A وجود داشته‌باشد به گونه‌ای که برای هر $\epsilon, \delta \in (۰,۱)$ و برای هر $h \in H$ اگر مجموعه آموزشی با اندازه $m > m_H^{NUL}(\epsilon, \delta, h)$ به صورت مستقل از توزیع $\mathcal{D}$ از روی فضای نمونه، نمونه برداری شده باشند. آنگاه با احتمال دست کم $۱ - \delta$ نابرابری زیر برای همه فرضیه‌های درون H برقرار است.

$$\mathbf{R}(A(S)) \leq \mathbf{R}(h) + \epsilon. \tag{۲.۴}$$

هرچه فضای H بزرگتر باشد فرضیه‌های این فضا انعطاف پذیرتر و درنتیجه هزینه این انعطاف پذیری بزرگ شدن مجموعه آموزشی برای یادگیری آن خواهد شد. در یادگیری نایکنواخت فضای فرضیه گروه‌بندی می‌شود بطوریکه هرچه یک گروه پیچیده‌تر باشد پیچیدگی نمونه‌ای یادگیری از آن گروه بزرگتر خواهدشد.

قضیه ۱.۴ (یادگیری نا یکنواخت).
کلاس فرضیه H شامل دسته بندهای $h : \mathcal{X} \mapsto \{۰,۱\}$ قابلیت یادگیری نا یکنواخت را دارد اگر و تنها اگر از اجتماع شمارای کلاس فرضیه‌هایی که قابلیت یادگیری احتمالا تقریبا درست بدون پیش فرض را داشته باشند تشکیل شده باشد.

اثبات قضیه بالا از قضیه زیر نتیجه گیری می شود.

قضیه ۲.۴ (یادگیر نا یکنواخت).
اگر $H = \bigcup_{n \in \mathbb{N}} H_n$ باشد و هر کدام از کلاس های فرضیه H_n دارای ویژگی همگرایی یکنواخت باشند آنگاه کلاس فرضیه H یادگیر نایکنواخت است.

اثبات قضیه ۱.۴ برای اثبات بخش اگر، فرض کنیدکه H به صورت $H = \bigcup_{n\in\mathbb{N}} H_n$ تعریف شده باشد که هر کدام از H_n ها دارای قابلیت یادگیری احتمالا تقریبا درست بدون پیش فرض باشد. از قضیه مهم یادگیری داشتیم که هر کدام از H_n ها دارای ویژگی همگرایی یکنواخت می‌باشند. بر اساس قضیه ۲.۴ نتیجه می گیریم که H دارای ویژگی یادگیری نا یکنواخت است.

برای اثبات بخش تنها اگر، فرض کنید که H با استفاده از الگوریتم A یادگیر نایکنواخت باشد. انگاه برای هر $n \in \mathbb{N}$، فرض کنید که H_n را به صورت زیر تعریف کنیم.

$$H_n = \left\{ h \in H \;\middle|\; m_H^{NUL}\left(\frac{۱}{۸}, \frac{۱}{۷}, h\right) \leq n \right\}. \tag{۳.۴}$$

روشن است که $H = \bigcup_{n\in\mathbb{N}} H_n$. به علاوه بر اساس تعریف m_H^{NUL}، می دانیم که برای هر توزیع $\mathcal{D}$ که شرط تحقق پذیری [1] را براساس H_n ارضا نماید با احتمال دست کم $\frac{۶}{۷}$ روی $S \sim \mathcal{D}^m$ داریم $\mathbf{R}(A(S)) \leq \frac{۱}{۸}$. براساس قضیه اساسی یادگیری، این نتیجه می دهد که بعد $VC(H_n)$ باید متناهی باشد و در نتیجه H_n قابلیت یادگیری احتمالا تقریبا درست بدون پیش فرض را دارد. ■

مثال زیر نشان می دهد که یادگیری نایکنواخت نسخه راحت شده یادگیری احتمالا تقریبا درست بدون پیش فرض است. یعنی فضاهای فرضیه‌ای وجود دارند که قابلیت یادگیری نایکنواخت را دارند اماقابلیت یادگیری احتمالا تقریبا درست بدون پیش فرض را ندارند.

مثال ۱.۴ (قابلیت یادگیری نایکنواخت خانواده چندجمله‌ای ها).
مساله دسته بندی دودویی با فضای نمونه $X = \mathbb{R}$ را در نظر بگیرید که برای هر $n \in \mathbb{N}$، فرضیه H_n چندجمله‌ای از درجه n باشد یعنی

$$H_n = \{h_p \in H \mid h_p(x) = \mathsf{sgn}\left(P_n(x)\right)\}. \tag{۴.۴}$$

که $P_n(x)$ چندجمله ای از درجه n است. فرض کنید که فضای فرضیه به صورت $H = \bigcup_{n\in\mathbb{N}} H_n$ باشد. آنگاه H مجموعه تمام چندجمله ای های ممکن است. به راحتی می توان نشان داد که $VC(H_n) = n + ۱$ و $VC(H) = \infty$ است. در نتیجه H دارای قابلیت یادگیری احتمالا تقریبا درست بدون پیش فرض نیست اما براساس قضیه ۱.۴ مجموعه فرضیه های H دارای قابلیت یادگیری نایکنواخت را دارد.

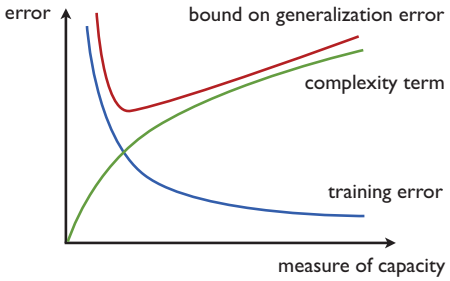
حال این پرسش مطرح است که آیا الگوریتمی برای یادگیری نایکنواخت وجود دارد؟ در ادامه به بررسی چنین الگوریتمی می‌پردازیم.

۲.۴ کمینه سازی هزینه ساختاری

پیش از این دیدیم که با کمک دانش پیشین می توانیم کلاس فرضیه ها را پیدا کنیم یعنی در واقع امیدواریم که H دارای فرضیه‌ای باشد که برای مساله داده شده دارای خطای کمی باشد. یک روش دیگر برای به کاربردن دانش پیشین، مشخص نمودن ترجیح در کلاس فرضیه های H است که در کمینه سازی هزینه ساختاری به کار می رود. روش کمینه سازی هزینه ساختاری، دنباله ای از فرضیه ها را به صورت افزایشی اندازه آنها به صورت زیر در نظر می گیرید.

$$H_۰ \subseteq H_۱ \subseteq H_۲ \subseteq \ldots \subseteq H_n \subseteq \ldots \tag{۵.۴}$$

[1] Realizability



شکل ۱.۴ کمینه سازی هزینه ساختاری

و تلاش می کند پاسخ خطای تجربی کمینه برای هر H_n را پیدا نماید و سپس از این کلاس‌های مختلف یک فرضیه را پیدا و به عنوان خروجی اعلام نماید. فرضیه انتخاب شده فرضیه ای است که تابع هدف زیر را که به صورت مجموع خطای تجربی و پیچیدگی فرضیه تعریف می شود کمینه باشد.

$$h_m^{SRM} = \underset{h\in H_n, n\in\mathbb{N}}{\text{argmin}} \left[\hat{\mathbf{R}}(h) + Complexity(H_n, m)\right] \tag{۶.۴}$$

که تابع $Complexity(H_n, m)$ به قدرت توصیف فضای فرضیه که ظرفیت یا غنای فضای فرضیه را مشخص می نماید و اندازه مجموعه آموزشی وابسته است. شکل ۱.۴ خطای کمینه سازی هزینه ساختاری را نشان می دهد.

علیرغم اینکه روش کمینه سازی هزینه ساختاری از پشتوانه نظری بسیار غنی برخوردار است اما دارای هزینه محاسباتی بسیار زیادی است و می‌بایست الگوریتم کمینه سازی هزینه تجربی را برای هر خانواده از H_n اجرا نماید. آیا تعداد مسایلی که باید توسط الگوریتم کمینه سازی خطا حل نماییم بی نهایت است؟ این تعداد مسایل بینهایت نیست زیرا اگر برای یک $n > n_\circ$ خطای تجربی برابر صفر باشد دیگر نیاز به حل آن مسایل نیست و در نتیجه تعداد این مسایل بی نهایت نیست.

همان گونه که پیش از این بیان شد، از دانش پیشین می‌توانیم برای مشخص نمودن ترجیح در کلاس فرضیه های H استفاده نماییم. در کمینه سازی هزینه ساختاری، در ابتدا فرض می‌کنیم که H را می‌توان به صورت $H = \bigcup_{n\in\mathbb{N}} H_n$ نوشت و سپس تابع وزنی را مشخص نمود که $w : \mathbb{N} \mapsto [\circ, ۱]$ به هر کلاس فرضیه H_n داده می‌شود به گونه‌ای که مقدار بیشتر وزن به معنای ترجیح بیشتر است. برای هر $n \in \mathbb{N}$ فرض می‌کنیم که H_n دارای ویژگی همگرایی یکنواخت است و پیچیدگی نمونه‌ای آن برابر $m_{H_n}^{UC}(\epsilon, \delta)$ است. در ادامه بیان می‌کنیم که چگونه می‌توانیم چنین دانش پیشینی را برای یادگیری به کار بریم. اگر تابع $\epsilon_n : \mathbb{N} \times (\circ, ۱) \mapsto (\circ, ۱)$ را به صورت زیر تعریف نماییم

$$\epsilon_n(m, \delta) = \min\{\epsilon \mid m_H^{UC}(\epsilon, \delta) \le m\}. \tag{۷.۴}$$

یعنی برای هر مجموعه آموزشی با طول ثابت m، ما به کمترین کران بین خطای واقعی و خطای تجربی علاقه‌مند هستیم. از تعریف همگرایی یکنواخت و ϵ_n می‌توان نتیجه گرفت که برای هر m و δ با احتمال دست کم $۱ - \delta$ روی انتخاب $S \sim \mathcal{D}^m$ نابرابری زیر برای همه $h \in H_n$ برقرار است.

$$|\mathbf{R}(h) - \hat{\mathbf{R}}(h)| \le \epsilon_n(m, \delta). \tag{۸.۴}$$

فرض کنید که $w : \mathbb{N} \mapsto [\circ, ۱]$ تابع وزن روی کلاس فرضیه ها باشد به گونه ای که $\sum_{n=۱}^{\infty} w(n) \le ۱$. این تابع وزن اهمیت یا پیچیدگی کلاس فرضیه را نشان می دهد. اگر H مجموعه متناهی از M فضای فرضیه $H_۱, H_۲, \ldots, H_M$ باشد انگاه می توانیم به هر کدام از کلاس

ها وزن $\frac{۱}{M}$ تخصیص بدهیم. این وزن یکنواخت نشان دهنده نداشتن دانش پیشین (ترجیح) است. اما اگر H از اجتماع شمارای کلاس فرضیه های H_n تشکیل شده باشد. آنگاه وزن یکنواخت امکان پذیر نیست اما روش های دیگر وزن دهی امکان پذیر است. برای نمونه توابع زیر دو نمونه از توابع وزن ممکن هستند.

$$w(n) = \frac{۶}{(\pi n)^۲} \quad (۹.۴)$$

$$w(n) = ۲^{-n} \quad (۱۰.۴)$$

الگوریتم کمینه‌سازی ساختاری تلاش می‌کند کران را کمینه نماید. یعنی هدف پیدا کردن فرضیه‌ای است که کران بالای خطای واقعی را کمینه نماید. در ادامه روش سرراستی را جهت تعریف تابع وزن معرفی می نماییم.

قضیه ۳.۴ (کران خطا برای کمینه سازی هزینه ساختاری).
فرض کنید $w : \mathbb{N} \mapsto [۰, ۱]$ یک تابع وزن با شد به گونه‌ای که $\sum_{n=۱}^{\infty} w(n) \leq ۱$ و همچنین فرض کنید H روی کلاس فرضیه‌ای باشد که بتوان به صورت $H = \bigcup_{n \in \mathbb{N}} H_n$ نوشت که هر کدام از H_n ها دارای ویژگی همگرایی یکنواخت با پیچیدگی نمونه‌ای $m^{UC}_{H_n}$ باشد اگر ϵ_n به صورت $\epsilon_n(m, \delta) = \min\{\epsilon \mid m^{UC}_{H_n}(\epsilon, \delta) \leq m\}$ تعریف شده باشد. آنگاه برای هر $\epsilon, \delta \in (۰, ۱)$ و هر توزیع $\mathcal{D}$ با احتمال دست کم $۱ - \delta$ برای مجموعه آموزشی $S \sim \mathcal{D}^m$ نابرابری زیر برای همه $n \in \mathbb{N}$ و همه $h \in H_n$ برقرار است.

$$|\mathbf{R}(h) - \hat{\mathbf{R}}(h)| \leq \epsilon_n(m, w(n) \times \delta) \quad (۱۱.۴)$$

و به بیانی دیگر برای هر $\delta \in (۰, ۱)$ و توزیع $\mathcal{D}$، با احتمال دست کم $۱ - \delta$ نابرابری زیر برای همه $h \in H$ برقرار است.

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}(h) + \min_{h \in H_n, n \in \mathbb{N}} \epsilon_n(m, w(n) \times \delta) \quad (۱۲.۴)$$

برهان برای هر n مقدار δ_n را به صورت $\delta_n = w(n) \times \delta$ تعریف می کنیم. با توجه به اینکه ویژگی همگرایی یکنواخت برای همه nها مطابق رابطه (۸.۴) برقرار است. اگر n را ثابت در نظر گیریم لذا با احتمال دست کم $۱ - \delta$ روی انتخاب $S \sim \mathcal{D}^m$ نابرابری زیر برای همه $h \in H_n$ برقرار است.

$$|\mathbf{R}(h) - \hat{\mathbf{R}}(h)| \leq \epsilon_n(m, \delta_n) \quad (۱۳.۴)$$

با بکارگیری کران اجتماعات [۲] روی $n = ۱, ۲, ۳, \ldots$ با احتمال دست کم $۱ - \delta \sum_n w(n) \geq ۱ - \delta$ نابرابری زیر برای همه nها و همه $h \in H_n$ها برقرار است.

$$|\mathbf{R}(h) - \hat{\mathbf{R}}(h)| \leq \epsilon_n(m, \delta_n) \quad (۱۴.۴)$$

و بدین ترتیب اثبات قضیه کامل می شود. ■

اگر $n(h)$ را به صورت زیر تعریف کنیم

$$n(h) = \min\{n \mid h \in H_n\} \quad (۱۵.۴)$$

[۲] Union bound

Structural Risk Minimization (SRM)

prior knowledge:
$\mathcal{H} = \bigcup_n \mathcal{H}_n$ where $\mathcal{H}_n$ has uniform convergence with $m^{\text{UC}}_{\mathcal{H}_n}$
$w : \mathbb{N} \to [0,1]$ where $\sum_n w(n) \leq 1$
define: ϵ_n as in Equation (7.1); $n(h)$ as in Equation (7.4)
input: training set $S \sim \mathcal{D}^m$, confidence δ
output: $h \in \text{argmin}_{h \in \mathcal{H}} \left[L_S(h) + \epsilon_{n(h)}(m, w(n(h)) \cdot \delta) \right]$

شکل ۲.۴ الگوریتم کمینه سازی هزینه ساختاری

آنگاه رابطه (۱۲.۴) نتیجه زیر را می دهد.

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}(h) + \epsilon_{n(h)}(m, w(n(h)) \times \delta) \tag{۱۶.۴}$$

روش کمینه سازی هزینه ساختاری فضا را برای پیداکردن فرضیه ای که کران بالا را کمینه نماید جستجو می نماید و این روش در الگوریتم ۲.۴ ارایه شده است.

همچنین می توان نشان داد که روش کمینه سازی هزینه ساختاری می تواند برای یادگیری نایکنواخت هر فضای فرضیه‌ای که از اجتماع شمارای فضای فرضیه هایی که ویژگی همگرایی یکنواخت را دارند به کار رود.

قضیه ۴.۴ (قابلیت یادگیری نایکنواخت با روش کمینه‌سازی هزینه ساختاری).
فرض کنید $H = \bigcup_{n \in N} H_n$ بطوریکه H_n دارای ویژگی همگرایی یکنواخت با پیچیدگی نمونه‌ای $m^{UC}_{H_n}$ باشد. همچنین تابع وزن $w : N \mapsto [۰, ۱]$ به صورت $w(n) = \frac{۶}{\pi^۲ n^۲}$ تعریف شده باشد. آنگاه H قابلیت یادگیری نایکنواخت توسط روش کمینه سازی هزینه ساختاری با نرخ زیر را دارد.

$$m^{NUL}_H(\epsilon, \delta, h) \leq m^{UC}_{H_{n(h)}}\left(\frac{\epsilon}{۲}, \frac{۶\delta}{(\pi n(h))^۲}\right) \tag{۱۷.۴}$$

برهان فرض کنید الگوریتم A، یک الگوریتم کمینه سازی هزینه ساختاری با تابع وزن $w(n)$ باشد. آنگاه برای هر $h \in H$، ϵ و δ فرض کنید نابرابری $m \geq m^{UC}_{H_{n(h)}}\left(\frac{\epsilon}{۲}, w(n(h))\delta\right)$ برقرار باشد. از آنجایی که $\sum_n w(n) = ۱$، با استفاده از قضیه ۳.۴، با احتمال دست کم $۱ - \delta$ برای مجموعه آموزشی $S \sim \mathcal{D}^m$ نابرابری زیر برای همه فرضیه های $h \in H$ برقرار است.

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}(h) + \epsilon_{n(h)}(m, w(n(h))\delta). \tag{۱۸.۴}$$

نابرابری بالا برای فرضیه‌ي $A(S)$ که توسط الگوریتم کمینه سازی هزینه ساختاری تولید می شود برقرار است. از تعریف کمینه سازی هزینه ساختاری می توانیم نابرابری زیر را بدست آوریم.

$$\mathbf{R}(A(S)) \leq \min_h \hat{\mathbf{R}}(h') + \epsilon_{n(h')}(m, w(n(h'))\delta) \tag{۱۹.۴}$$

$$\leq \hat{\mathbf{R}}(h) + \epsilon_{n(h)}(m, w(n(h))\delta) \tag{۲۰.۴}$$

در پایان اگر $m \geq m^{UC}_{H_{n(h)}}(\epsilon/۲, w(n(h))\delta)$ باشد نابرابری $\epsilon_{n(h)}(m, w(n(h))\delta) \leq \epsilon/۲$ هم برقرار است. از طرفی ویژگی همگرایی یکنواخت، هر H_n با احتمال دست کم $۱ - \delta$ نابرابری $\hat{\mathbf{R}}(h) \leq \mathbf{R}(h) + \epsilon/۲$ برقرار است با ترکیب این نابرابری با نابرابری (۲۰.۴)،

نابرابری زیر بدست می آید

$$\mathbf{R}(A(S)) \quad \leq \quad \mathbf{R}(h) + \epsilon. \tag{۲۱.۴}$$

که اثبات قضیه تکمیل می گردد. ■

پیش از این نشان دادیم که هر فضای فرضیه که از اجتماع شمارای مجموعه فرضیه هایی با بعد VC متناهی تشکیل شده باشند قابلیت یادگیری نایکنواخت را دارد. می توان نشان داد که برای هر فضای نامتناهی $\mathcal{X}$، مجموعه همه دسته بند های دو دسته‌ای روی فضای $\mathcal{X}$ را نمی توان اجتماع شمارای مجموعه فرضیه هایی با بعد VC متناهی تشکیل داد. این بدان معنا است که قضیه No free lunch برای یادگیری نایکنواخت نیز برقرار است. هنگامی که فضای نمونه نامتناهی است هیچ یادگیر نایکنواختی برای مجموعه همه دسته بندهای ممکن دو دسته‌ای وجود ندارد. هرچند برای هریک از این دسته بندها، یک الگوریتم یادگیری وجود دارد.

نکته دیگری که در یادگیری نایکنواخت وجود دارد این است که نقش دانش پیشین برای یادگیری نایکنواخت چیست و آیا این دانش پیشین سبب افزایش پیچیدگی نمونه‌ای می‌شود یا کاهش آن؟ در قابلیت یادگیری یکنواخت، معمولا دانش پیشین به صورت انتخاب فضای فرضیه در الگوریتم یادگیری وارد می‌شود در حالیکه در یادگیری نایکنواخت این دانش پیشین به صورت ترجیح یا تابع وزن در الگوریتم یادگیری وارد می‌شود. بدیهی است این نوع دانش پیشین ضعیف‌تر از گونه ای است که در قابلیت یادگیری یکنواخت است و هزینه این ضعیف کردن دانش پیشین، افزایش پیچیدگی نمونه‌ای هر کدام از فضاهای H_n است. اگر دانش پیشین در باره n داشته باشیم می توانیم الگوریتم کمینه‌سازی خطای تجربی را روی فضای H_n اجرا نماییم که در این حالت پیچیدگی نمونه‌ای برابر است با $C\frac{n+\log\left(\frac{۱}{\delta}\right)}{\epsilon^۲}$ که $VC(H_n) = n$ و C یک مقدار ثابت است. اگر دانش‌پیشینی در باره n نداشته باشیم الگوریتم کمینه سازی هزینه ساختاری را روی فضای H اجرا نموده و در این حالت با استفاده از قضیه ۴.۴ پیچیدگی نمونه ای یادگیر نایکنواخت برابر است با

$$m_H^{NUL}(\epsilon, \delta, h) \leq m_{H_n}^{UC}(\epsilon/۲, w(n)\delta) = O\left(\frac{n + \log\left(\frac{۱}{w(n)}\right) + \log\left(\frac{۱}{\delta}\right)}{\epsilon^۲}\right) \tag{۲۲.۴}$$

برای نمونه اگر تابع وزن $w(n) = \frac{۱}{۲n^۲}$ در نظر گرفته شود و تفاضل پیچیدگی نمونه ای را محاسبه نماییم خواهیم داشت.

$$\begin{aligned} m_H^{NUL}(\epsilon, \delta, h) - m_{H_n}^{UC}(\epsilon/۲, w(n)\delta) &= O\left(\frac{n + \log\left(\frac{۱}{w(n)\delta}\right)}{\epsilon^۲}\right) - O\left(\frac{n + \log\left(\frac{۱}{\delta}\right)}{\epsilon^۲}\right) \\ &= O\left(\frac{\log\left(\frac{۱}{w(n)}\right)}{\epsilon^۲}\right) \qquad (۲۳.۴) \\ &= O\left(\frac{\log n}{\epsilon^۲}\right) \qquad (۲۴.۴) \end{aligned}$$

در نتیجه هزینه ضعیف کردن دانش پیشین افزایش پیچیدگی نمونه‌ای یادگیر است. در نتیجه برای یک مساله مشخص، یادگیر نایکنواخت به تعداد بیشتری نمونه نیاز دارد تا یادگیر یکنواخت. حال فرض کنید که $w : H \mapsto [۰, ۱]$ تابع وزن به صورت $\sum_{h\in H} w(h) \leq ۱$ باشد. در قضیه نشان می‌دهیم که یادگیری فضای فرضیه H که وزن آن به صورت بالا تعریف شده باشد قابلیت یادگیری نایکنواخت را دارد.

قضیه ۵.۴ ().
اگر H فضای فرضیه ها و $w : H \mapsto [۰, ۱]$ تابع وزن به صورت $\sum_{h\in H} w(h) \leq ۱$ باشد. آنگاه برای هر $h \in H$ و هر $\delta > ۰$

و هر توزیع $\mathcal{D}$ و برای هر مجموعه آموزشی S با m عضو نابرابری زیر برقرار است.

$$\underset{S\sim\mathcal{D}^m}{\mathbb{P}}\left[\mathbf{R}(h) \le \hat{\mathbf{R}}(h) + \sqrt{\frac{\ln\left(\frac{۱}{w(h)}\right) + \ln(۲/\delta)}{۲m}}\right] \ge ۱ - \delta. \tag{۲۵.۴}$$

برهان بطور معادل نابرابری زیر را اثبات می کنیم.

$$\underset{S\sim\mathcal{D}^m}{\mathbb{P}}\left[\exists h \in H \;\middle|\; |\hat{\mathbf{R}}(h) - \mathbf{R}(h)| \ge \sqrt{\frac{\ln\left(\frac{۱}{w(h)}\right) + \ln(۲/\delta)}{۲m}}\right] \le \delta. \tag{۲۶.۴}$$

اگر برای هر $h \in H$ مقدار ϵ_h را به صورت زیر تعریف کنیم.

$$\epsilon_h = \sqrt{\frac{\ln\left(\frac{۱}{w(h)}\right) + \ln(۲/\delta)}{۲m}}. \tag{۲۷.۴}$$

با استفاده از قاعده کران اجتماعات خواهیم داشت

$$\underset{S\sim\mathcal{D}^m}{\mathbb{P}}\left[\exists h \in H \;\middle|\; |\hat{\mathbf{R}}(h) - \mathbf{R}(h)| \ge \epsilon_h\right] \le \sum_{h\in H} \underset{S\sim\mathcal{D}^m}{\mathbb{P}}\left[|\hat{\mathbf{R}}(h) - \mathbf{R}(h)| \ge \epsilon_h \;\middle|\; .\right] \tag{۲۸.۴}$$

با استفاده از نابرابری هافدینگ می توانیم کران بالا را محاسبه نماییم.

$$\begin{aligned}
\underset{S\sim\mathcal{D}^m}{\mathbb{P}}\left[|\hat{\mathbf{R}}(h) - \mathbf{R}(h)| \ge \epsilon_h\right] &\le ۲\exp\left[-۲m\epsilon_h^۲\right] \\
&= ۲\exp\left[-\ln\left(\frac{۲}{w(h)\delta}\right)\right] \\
&= ۲\exp\left[\ln\left(\frac{w(h)\delta}{۲}\right)\right] \\
&= w(h)\delta \\
&\le \delta. \qquad (۲۹.۴)
\end{aligned}$$

با جایگزینی نابرابری بالا در نابرابری (۲۸.۴) خواهیم داشت.

$$\begin{aligned}
\underset{S\sim\mathcal{D}^m}{\mathbb{P}}\left[\exists h \in H \;\middle|\; |\hat{\mathbf{R}}(h) - \mathbf{R}(h)| \ge \sqrt{\frac{\ln\left(\frac{۱}{w(h)}\right) + \ln(۲/\delta)}{۲m}}\right] &\le \sum_{h\in H} \underset{S\sim\mathcal{D}^m}{\mathbb{P}}\left[|\hat{\mathbf{R}}(h) - \mathbf{R}(h)| \ge \epsilon_h \;\middle|\; \right] \\
&\le \delta \sum_{h\in H} w(h) \qquad (۳۰.۴) \\
&\le \delta. \qquad (۳۱.۴)
\end{aligned}$$

نابرابری آخر با توجه به قید $\sum_{h\in H} w(h) \le ۱$ نتیجه گرفته شده است. ■

قضیه ۶.۴ ().
فرض کنید H فضای فرضیه و $w : H \mapsto [۰, ۱]$ تابع وزن به صورت $\sum_{h\in H} w(h) \leq ۱$ باشد. اگر الگوریتم یادگیری A با دریافت مجموعه آموزشی S، فرضیه

$$\hat{h} = \underset{h\in H}{\text{argmin}}\, \hat{\mathbf{R}}(h) + \sqrt{\frac{\ln\left(\frac{۱}{w(h)}\right) + \ln(۲/\delta)}{۲m}} \tag{۳۲.۴}$$

را تولید نماید آنگاه برای هر $h \in H$ و هر $\delta > ۰$ و هر توزیع $\mathcal{D}$، اگر مجموعه آموزشی دست کم $m \geq ۲\frac{\ln\left(\frac{۱}{w(h)}\right)+\ln(۲/\delta)}{\epsilon^۲}$ نمونه داشته باشد. آنگاه نابرابری زیر برقرار است.

$$\underset{S\sim\mathcal{D}^m}{\mathbb{P}}\left[\mathbf{R}(\hat{h}) \geq \mathbf{R}(h) + \epsilon\right] \leq \delta. \tag{۳۳.۴}$$

برهان از قضیه ۵.۴ نتیجه می گیریم که با احتمال دست کم $\delta - ۱$ دو نابرابری زیر برقرار هستند.

$$\mathbf{R}(\hat{h}) \leq \hat{\mathbf{R}}(\hat{h}) + \sqrt{\frac{\ln\left(\frac{۱}{w(h)}\right) + \ln(۲/\delta)}{۲m}} \tag{۳۴.۴}$$

$$\hat{\mathbf{R}}(h) \leq \mathbf{R}(h) + \sqrt{\frac{\ln\left(\frac{۱}{w(h)}\right) + \ln(۲/\delta)}{۲m}} \tag{۳۵.۴}$$

اگر m را از صورت قضیه در رابطه بالا جایگزین نماییم خواهیم داشت.

$$\mathbf{R}(\hat{h}) \leq \hat{\mathbf{R}}(h) + \frac{\epsilon}{۲}. \tag{۳۶.۴}$$

با توجه به اینکه $\hat{h}$ کران قید شده در صورت قضیه را کمینه می‌کند خواهیم داشت.

$$\mathbf{R}(\hat{h}) \leq \hat{\mathbf{R}}(\hat{h}) + \sqrt{\frac{\ln\left(\frac{۱}{w(h)}\right) + \ln(۲/\delta)}{۲m}} \tag{۳۷.۴}$$

$$\hat{\mathbf{R}}(h) \leq \mathbf{R}(h) + \sqrt{\frac{\ln\left(\frac{۱}{w(h)}\right) + \ln(۲/\delta)}{۲m}} \tag{۳۸.۴}$$

$$\leq \hat{\mathbf{R}}(h) + \frac{\epsilon}{۲} \tag{۳۹.۴}$$

$$\leq \mathbf{R}(h) + \frac{\epsilon}{۲} + \frac{\epsilon}{۲} \tag{۴۰.۴}$$

$$= \mathbf{R}(h) + \epsilon. \tag{۴۱.۴}$$

که اثبات قضیه کامل می‌گردد. ■

۳.۴ کمینه طول توصیف

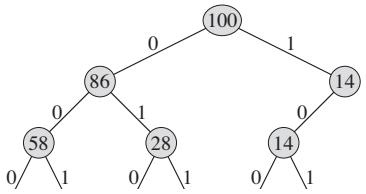
فرض کنید که H فضای شمارایی از فرضیه ها باشد در این صورت می توانیم آن را به صورت $H = \bigcup_{n\in\mathbb{N}}\{h_n\}$ بنویسیم. پیش از این با کمک نابرابری هافدینگ نشان دادیم که یک فضای تک فرضیه‌ای دارای ویژگی همگرایی یکنواخت با نرخ $m^{UC}(\epsilon,\delta) = \frac{\log(۲/\delta)}{۲\epsilon^۲}$ است. در نتیجه مقدار ϵ_n برابر است با $\epsilon_n(\epsilon,\delta) = \sqrt{\frac{\log(۲/\delta)}{۲\epsilon^۲}}$ و قاعده کمینه سازی هزینه ساختاری به صورت زیر است.

$$\underset{h_n\in H}{\text{argmin}} \left[\hat{\mathbf{R}}(h_n) + \sqrt{\frac{\log(۱/w(n)) + \log(۲/\delta)}{۲m}}\right] \tag{۴۲.۴}$$

به طور معادل می توانیم تابع وزن را به صورت $w : H \mapsto [۰,۱]$ در نظر گیریم و قاعده کمینه سازی هزینه ساختاری به صورت زیر می‌شود.

$$\underset{h\in H}{\text{argmin}} \left[\hat{\mathbf{R}}(h_n) + \sqrt{\frac{-\log w(h) + \log(۲/\delta)}{۲m}}\right] \tag{۴۳.۴}$$

در این حالت دانش پیشین به وسیله $w(h)$ که به هر فرضیه انتساب داده می شود مشخص می گردد. در این حالت به فرضیه هایی که فکر می‌کنیم بهتر هستند وزن بیشتری داده می شود و الگوریتم یادگیری به فرضیه هایی که وزن بیشتری دارند ترجیح بیشتری می‌دهد. در این بخش، یک تابع وزن روی H تعریف می کنیم که به طول توصیف فرضیه ها وابسته است. حال چگونه یک فرضیه را توصیف می کنیم؟ اگر $\Sigma = \{۰,۱\}$ مجموعه الفبا و Σ^* همه رشته های ممکن باشد. یک زبان برای توصیف فضای فرضیه H تابعی به صورت $\mathcal{L} : H \mapsto \Sigma^*$ است که هر عضو $h \in H$ را به یک توصیف $\mathcal{L}(h)$ نگاشت می کند و $\mathcal{L}(h)$ توصیف h نامیده می شود. درادامه طول این توصیف را با $|h|$ نشان می دهیم. زبان توصیف h باید ویژگی پیشوند_آزاد [۳] را داشته باشد. برای نمونه درخت هافمن که در شکل زیر نشان داده شده است یک زبان پیشوند_آزاد است.



شکل ۳.۴ درخت هافمن به عنوان نمونه ای از زبان پیشوند_آزاد.

هر مجموعه از رشته های پیشوند_آزاد دارای ویژگی زیر است که به نابرابری کرافت [۴] معروف می باشد.

لم ۱.۴ (نابرابری کرافت).
اگر $s \subset \{۰,۱\}^*$ یک مجموعه از رشته های پیشوند_آزاد باشند آنگاه نابرابری زیر برقرار است.

$$\sum_{\sigma\in s} \frac{۱}{۲^{|\sigma|}} \leq ۱. \tag{۴۴.۴}$$

برهان برای اثبات این لم، یک سکه سالم را به کار می‌بریم. یک رشته با طول صفر در ابتدا در نظر می‌گیریم. سکه را پرتاپ می‌کنیم اگر شیر آمد نماد ۱ و اگر خط آمد نماد ۰ را به رشته اضافه می‌کنیم. با افزودن هر بیت به رشته بررسی می‌کنیم که آیا رشته عضوی از S است یا نه. اگر عضوی از S باشد کار تولید این رشته تمام شده است و رشته‌ای دیگر را آغاز می‌کنیم. از آنجا که احتمال تولید یک رشته σ برابر است با $۲^{|\sigma|}$ و رشته‌ها مستقل از هم تولید شده و برخی از رشته‌ها نیز عضو S نیستند. بنابراین لم اثبات می‌گردد. ■

[۳] Prefix-free
[۴] Kraft inequality

با توجه به نابرابری کرافت، می توانیم تابع وزن $w(h)$ را به صورت زیر تعریف کنیم.

$$w(h) = \frac{۱}{۲^{|h|}}. \tag{۴۵.۴}$$

با توجه به تعریف بالا، می توانیم قضیه زیر را داشته باشیم.

قضیه ۷.۴ ().
فضای فرضیه H و زبان توصیف پیشوند_آزاد $\mathcal{L} : H \mapsto \{۰,۱\}^*$ برای توصیف این فضا را در نظر بگیرید. آنگاه برای هر مجموعه آموزشی $S \sim \mathcal{D}^m$، هر توزیع $\mathcal{D}$ و هر $\delta > ۰$ با احتمال دست کم $۱ - \delta$ نابرابری زیر برای همه $h \in H$ برقرار است.

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}(h) + \sqrt{\frac{|h| + \ln(۲/\delta)}{۲m}} \tag{۴۶.۴}$$

که $|h|$ طول توصیف $\mathcal{L}(h)$ است.

برهان اگر $w(h) = \frac{۱}{۲^{|h|}}$ تابع وزن باشد و با استفاده از قضیه ۳.۴ و جایگزینی $\epsilon_n(m,\delta) = \sqrt{\frac{\ln(۲/\delta)}{۲m}}$، و استفاده از نابرابری $|h| \ln ۲ < |h|$ خواهیم داشت

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}(h) + \sqrt{\frac{|h| \ln ۲ + \ln(۲/\delta)}{۲m}} \tag{۴۷.۴}$$

$$\leq \hat{\mathbf{R}}(h) + \sqrt{\frac{|h| + \ln(۲/\delta)}{۲m}}. \tag{۴۸.۴}$$

و بدین ترتیب قضیه اثبات می‌گردد. ■

روش یادگیری که کران $\hat{\mathbf{R}}(h) + \sqrt{\frac{|h|+\ln(۲/\delta)}{۲m}}$ را کمینه می کند روش کمینه سازی طول توصیف [۵] می گویند که الگوریتم آن در شکل زیر نشان داده شده است. در واقع این الگوریتم مصالحه‌ای بین خطای تجربی و طول توصیف برقرار می کند.

Minimum Description Length (MDL)

prior knowledge:
$\mathcal{H}$ is a countable hypothesis class
$\mathcal{H}$ is described by a prefix-free language over $\{0,1\}$
For every $h \in \mathcal{H}$, $|h|$ is the length of the representation of h
input: A training set $S \sim \mathcal{D}^m$, confidence δ
output: $h \in \text{argmin}_{h \in \mathcal{H}} \left[L_S(h) + \sqrt{\frac{|h|+\ln(2/\delta)}{2m}} \right]$

شکل ۴.۴ الگوریتم کمینه سازی طول توصیف

قضیه ۷.۴ بیان می دارد که دو فرضیه‌ای که خطای تجربی یکسانی دارند، فرضیه‌ای دارای کران خطای واقعی کمتری است که دارای طول توصیف کمتری باشد. نتیجه این قضیه با آنچه که اصل اوکم [۶] بیان می دارد که عبارت است از «توصیف کوتاهتر نسبت به توصیف طولانی تر ترجیح دارد» پیام یکسانی دارد. این قضیه همچنین بیان می دارد که فرضیه‌ای که طول توصیف طولانی‌تری دارد برای اینکه کران

[5] Minimum description length
[6] Occam razor

خطای واقعی که فرضیه با طول توصیف کمتری دارد برسد نیاز به مجموعه آموزشی با اندازه بزرگتری نیاز دارد. مساله‌ای که وجود دارد این است که طول توصیف یک فرضیه به زبان توصیف فرضیه وابسته است. اگر زبان توصیف فرضیه مستقل از مجموعه آموزشی و پیش از دیدن مجموعه آموزشی انتخاب شود آنگاه کران یاد شده در این قضیه برقرار است.

۴.۴ سازگاری

نمادهای یادگیری که تا کنون بررسی کردیم پیچیدگی نمونه‌ای را تابعی از ϵ, δ, h در نظر می‌گرفتند. سازگاری[7] یک قاعده یادگیری است که علاوه بر اینکه پیچیدگی نمونه‌ای را تابعی از ϵ, δ, h در نظر می‌گیرد تابعی از توزیع $\mathcal{P}$ که نمونه های آموزشی و آزمایشی بر اساس آن نمونه‌برداری شده‌اند در نظر می‌گیرد. سازگاری به صورت زیر تعریف می شود.

تعریف ۳.۴ **سازگاری** فرض کنید که $Z = \mathcal{X} \times \mathcal{Y}$ فضای دامنه، $\mathcal{P}$ مجموعه‌ای از توزیع‌ها روی Z و H فضای فرضیه باشد. یک قاعده یادگیری A نسبت به H و $\mathcal{P}$ سازگار است اگر یک تابع $m_H^{CON} : (۰,۱)^۲ \times H \times \mathcal{P} \mapsto \mathbb{N}$ وجود داشته باشد به طوریکه برای هر $\delta \in (۰,۱)$، هر $\epsilon \in (۰,۱)$، هر $h \in H$، هر $\mathcal{D} \in \mathcal{P}$، اگر مجموعه آموزشی با اندازه $m \geq m_H^{CON}(\epsilon, \delta, h, \mathcal{D})$، انتخاب شود آنگاه با احتمال دست کم $۱ - \delta$ روی مجموعه آموزشی $S \sim \mathcal{D}^m$ نابرابری زیر برقرار باشد

$$\mathbf{R}(A(S)) \leq \mathbf{R}(h) + \epsilon. \qquad (۴۹.۴)$$

اگر $\mathcal{P}$ تمام توزیع های ممکن باشد انگاه A را نسبت به H سازگار عمومی [8] می گویند. همانگونه که تعریف سازگاری نشان می دهد، سازگاری نسخه راحت شده یادگیری نایکنواخت است. یعنی همه یادگیرهای نایکنواخت، یادگیرهای سازگار هستند اما همه یادگیرهای سازگار، یادگیرهای نایکنواخت نیستند. برای نمونه، مثال زیر را در نظر بگیرید.

مثال ۲.۴ (یادگیرهای تنبل).
یک الگوریتم یادگیری تنبل را در نظر بگیرید که نمونه‌های آموزشی را ذخیره می کند و برچسب هر نمونه آزمون x با رای اکثریت روی برچسب نمونه‌های x که در مجموعه آموزشی وجود دارد تعیین می‌گردد و در صورتی که نمونه x در مجموعه آموزشی وجود نداشت یک برچسب پیش‌فرض انتخاب می‌گردد. می‌توان نشان داد برای هر دامنه شمارای $\mathcal{X}$ و مجموعه برچسب متناهی $\mathcal{Y}$ و هزینه صفر ـ یک الگوریتم تنبل یاد شده سازگار عمومی است اما یک یادگیر نایکنواخت نیست.

۵.۴ جمع بندی نمادهای مختلف یادگیری

تاکنون سه نماد مختلف یادگیری را بررسی نمودیم و اکنون به بررسی سودمندی این نمادها می‌پردازیم. سودمندی یک تعریف ریاضی به آنچه آن تعریف نیاز دارد وابسته است. در ادامه به بررسی سودمندی این تعاریف از نگاه اهداف زیر می‌پردازیم.

۱. هزینه فرضیه یادگرفته شده چیست؟

نخستین هدف پیداکردن کران‌ها، پیدا کردن خطای واقعی فرضیه یادگرفته شده است. روش های یادگیری احتمالا تقریبا درست و یادگیری نایکنواخت، کران خطای واقعی را براساس خطای تجربی پیدا می کنند. اما مدل سازگاری چنین کرانی را پیدا نمی کند. اما

[7] Consistency
[8] Universally consistent

همیشه می‌توان خطای واقعی فرضیه یادگرفته شده را با کمک مجموعه اعتبارسنجی تخمین زد.

۲. برای اینکه فرضیه یادگرفته شده به خوبی بهترین فرضیه باشد چند نمونه آموزشی مورد نیاز است؟

هنگامی که می‌خواهیم یک مساله یادگیری را حل نماییم نخستین پرسش این است که چند نمونه آموزشی جمع آوری نماییم؟ برای روش‌های یادگیری احتمالا تقریبا درست پاسخ روشن است اما برای روش‌های یادگیری نایکنواخت تعداد نمونه‌های آموزشی به بهترین فضای فرضیه و برای روش‌های یادگیری سازگار تعداد نمونه‌های آموزشی علاوه بر بهترین فضای فرضیه به توزیع نمونه‌برداری نیز وابسته است. از این سو مدل یادگیری احتمالا تقریبا درست تنها مدل قابل استفاده برای قابلیت یادگیری است. از سویی دیگر، هرچند ممکن است خطای براورد فرضیه یادگرفته شده کم باشد اما اگر خطای تقریب H بزرگ باشد ممکن است خطای واقعی آن زیاد باشد. برای پاسخ به این پرسش که «برای اینکه فرضیه یادگرفته شده به خوبی دسته بند بهینه بیز باشد چند نمونه آموزشی مورد نیاز است؟» برای این پرسش، مدل یادگیری احتمالا تقریبا درست پاسخ روشنی ندارد. این به این دلیل است که سودمندی روش‌های مبتنی بر مدل یادگیری احتمالا تقریبا درست به کیفیت دانش پیشین وابسته است.

هنگامی که فرضیه یادگرفته شده خطای واقعی زیادی داشته باشند، کران‌های بدست آمده برای مدل یادگیری احتمالا تقریبا درست کمک می کنند که چه کاری باید انجام دهیم. این بدان دلیل است که می‌توانیم کرانی از خطا که ریشه آن از بخش براورد خطا است را تعیین و در نتیجه مشخص نماییم ریشه چه مقدار از خطا از بخش تقریب است. اگر خطای تقریب بزرگ باشد باید فضای فرضیه را تغییر بدهیم. به طور مشابه اگر یک یادگیر نایکنواخت شکست بخورد می توانیم توابع وزن مختلف را در نظر بگیریم. در حالیکه اگر یک روش یادگیری سازگار شکست بخورد نمی دانیم که آیا بدلیل خطای براورد است یا خطای تقریب. اگر هم بدانیم که خطای براورد زیاد باشد نمی‌دانیم چند نمونه باید داشته باشیم تا خطای براورد کوچک داشته باشیم.

۳. چگونه یادبگیریم؟ چگونه دانش پیشین را نمایش دهیم؟

یکی از مهم‌ترین ویژگی های نظریه یادگیری این است که به پرسش «چگونه باید یادگرفت؟» پاسخ می‌دهد. تعریف یادگیری احتمالا تقریبا درست محدودیت یادگیری را به کمک قضیه No free lunch و همچنین لزوم دانش پیشین را مشخص می نماید. این مدل روشی روشن برای نمایش دانش پیشین با کمک انتخاب فضای فرضیه را ارایه می نماید پس از انتخاب فضای فرضیه، یک روش عمومی یادگیری که همان کمینه سازی خطای تجربی است را ارایه می نماید

مدل یادگیری نایکنواخت نیز روشی روشن برای نمایش دانش پیشین با کمک تابع وزن روی فضای فرضیه را ارایه می نماید پس از انتخاب تابع وزن، یک روش عمومی یادگیری که همان کمینه سازی خطای ساختاری است را ارایه می نماید. روش یادگیری کمینه سازی خطای ساختاری این برتری را دارد که در انتخاب مدل کمک می کند به ویژه زمانی که دانش پیشین جزیی در اختیار داشته باشیم.

اما مدل سازگاری بر عکس دو مدل پیشین، هیچ روشی برای نمایش دانش پیشین ارایه نمی‌کند. در بسیاری از حالت‌ها، نیازی به دانش پیشین نیست. برای نمونه، الگوریتم مثال ۲.۴ یک الگوریتم سازگار برای هر فضای نمونه شمارا با تعداد متناهی برچسب است. این اشاره می کند که سازگاری یک نیازمندی خیلی ضعیف است.

۴. باید کدام الگوریتم یادگیری را ترجیح دهیم؟

ممکن است دلیل آورده شود که با وجود اینکه سازگاری یک نیازمندی ضعیف است، خیلی خوب است که یک الگوریتم نسبت به مجموعه همه توابع از $\mathcal{X}$ به $\mathcal{Y}$ سازگار باشد. این روش تضمین می نماید که اگر به اندازه کافی نمونه آموزشی داشته باشیم دسته بند یادگرفته شده همیشه به خوبی دسته‌بند بهینه بیز خواهد بود. در نتیجه اگر دو الگوریتم که یکی سازگار باشد و دیگری سازگار نباشد الگوریتم سازگار را ترجیح می‌دهیم. هر چند این نتیجه‌گیری‌ی به دو دلیل مشکل‌ساز است. نخست آنکه در بسیاری از کاربردها و

برای بسیاری از توزیع ها، پیچیدگی نمونه‌ای الگوریتم های سازگار خیلی زیاد است و در عمل نمی‌توانیم این مقدار نمونه آموزشی جمع‌آوری نماییم. دوم اینکه، خیلی سخت نیست که الگوریتم‌های احتمالا تقریبا درست و نایکنواخت را نسبت به مجموعه همه توابع از $\mathcal{X}$ به $\mathcal{Y}$ به الگوریتم‌های سازگار تبدیل نماییم. از آنجایی که به سادگی می توان هر الگوریتم یادگیری را به الگوریتم سازگار تبدیل نمود بنابراین خردمندانه نیست که الگوریتم‌های سازگار را به الگوریتم های ناسازگار ترجیح دهیم.

پرسش‌ها

۱.۴ هدف یادگیری فضای فرضیه‌ای است به گونه‌ای که $\mathcal{X} = \mathbb{R}$ و به‌ازای تعداد متناهی نقطه برچسب $+۱$ و برای بقیه نقاط برچسب -۱ تولید نماید. آیا این فضا قابلیت یادگیری نایکنواخت را دارد؟ ادعای خود را اثبات نمایید.

۲.۴ آیا قضیه NFL برای یادگیری نایکنواخت برقرار است؟ ادعای خود را اثبات نمایید.

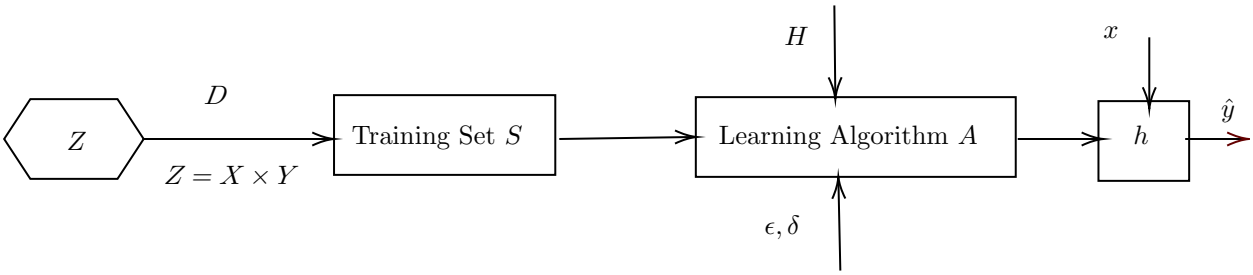
۵

پیچیدگی محاسباتی الگوریتم‌های یادگیری

در بخش های پیشین در باره ویژگی‌های آماری روش‌های یادگیری که عبارت‌است از تعداد نمونه‌های لازم برای یادگیری، سخن گفته شد و به بیانی دیگر، در باره اطلاعات مورد نیاز یادگیری سخن گفته‌شد. در کاربردهای عملی، هدف داشتن یادگیری خودکار است که در آن منابع محاسباتی بسیار مهم هستند. هنگامی که نمونه‌های آموزشی جمع‌آوری شد می‌بایست محاسباتی انجام گیرد تا فرضیه نهایی تولید شود. پس از تولید این فرضیه نیاز به محاسبات است تا برچسب نمونه‌های آزمون تعیین گردند. این منابع محاسباتی در کاربردهای عملی یادگیری ماشین بسیار مهم هستند. منابع لازم برای یادگیری دو دسته هستند. (۱) نمونه‌های برچسب خورده موردنیاز که توسط پیچیدگی نمونه‌ای مشخص می‌شوند و (۲) منابع محاسباتی لازم که توسط پیچیدگی محاسباتی مشخص می‌شوند. در این بخش پیچیدگی محاسباتی و به‌طور خاص زمان اجرای الگوریتم‌های یادگیری را بررسی می کنیم.

زمان اجرای یک الگوریتم به ماشین استفاده، زبان برنامه نویسی، الگوریتم طراحی شده و همچنین پیاده سازی آن وابسته است. برای جلوگیری از وابستگی زمان اجرا به ماشین (به طور مشابه برای عوامل دیگر)، زمان اجرای الگوریتم‌ها را به‌صورت مجانبی محاسبه می‌کنیم. برای نمونه زمان اجرای الگوریتم مرتب‌سازی ادغامی از مرتبه $O(n \log n)$ است که n اندازه ورودی است. این بدان معنا است که زمان اجرای این الگوریتم در هر ماشینی به‌صورت زیر تعریف می‌شود. «ثابت‌های c و n_0 وجود دارند که به ماشین وابسته هستند به گونه‌ای که برای هر مقدار $n > n_0$ زمان اجرای الگوریتم مرتب سازی ادغامی (به ثانیه) حداکثر $cn \log n$ است.» یک الگوریتم را کارا می گویند اگر در زمان از مرتبه $O(p(n))$ اجرا شود که $p(n)$ یک چند جمله‌ای براساس n است. این نوع تحلیل، زمان اجرای الگوریتم را بر اساس اندازه ورودی (n) محاسبه می نماید. برای نمونه در الگوریتم مرتب سازی ادغامی، تعداد اعضا آرایه ورودی به عنوان اندازه ورودی در نظر گرفته می‌شود.

در ادامه این فصل نخست پیچیدگی محاسباتی الگوریتم بطور رسمی تعریف می‌شود و سپس چندین مثال و شیوه محاسبه زمان اجرا



شکل ۱.۵ یک الگوریتم یادگیری نمونه

در آنها بررسی می‌گردد و همچنین نشان می دهیم که چگونه می‌توان با تغییر فضای فرضیه مسایل یادگیری را حل و در پایان سختی مسایل یادگیری بررسی می‌شود.

۱.۵ پیچیدگی محاسباتی یادگیری

یک الگوریتم یادگیری به فضای نمونه $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$، فضای فرضیه‌های H، تابع هزینه ℓ و نمونه‌های آموزشی که بر اساس توزیع ناشناخته $\mathcal{D}$ و به صورت مستقل از فضای نمونه $\mathcal{Z}$، نمونه برداری شده‌اند دسترسی دارد. برای هر مقدار از پارامترهای ϵ و δ الگوریتم یادگیری باید فرضیه hی را تولید نماید که با احتمال دست کم $1 - \delta$ نابرابری زیر برقرار باشد.

$$\mathbf{R}(h) \leq \hat{\mathbf{R}}(h) + \epsilon. \tag{۱.۵}$$

به بیانی دیگر یک الگوریتم یادگیری ماشین را می‌توان به‌صورت شکل ۱.۵ نشان داد.

برای تحلیل الگوریتم‌های یادگیری از روش‌های استاندارد تحلیل الگوریتم‌ها استفاده می‌کنیم که در آنها (۱) یک ماشین انتزاعی مانند ماشین تورینگ در نظر می‌گیریم و (۲) تحلیل مجانبی الگوریتم را انجام می دهیم. نخستین مساله‌ای که می‌بایست مشخص شود تعیین اندازه ورودی است. بر اساس اینکه اندازه ورودی را چه در نظر بگیریم تحلیل های متفاوتی خواهیم داشت. برای نمونه حالت های زیر را در نظر بگیرید.

۱. اندازه ورودی برابر با اندازه مجموعه آموزشی S باشد.

اندازه مجموعه آموزشی را نمی‌توان به تنهایی به عنوان اندازه ورودی در نظر گرفت چون اگر اندازه مجموعه آموزشی از $m(\epsilon, \delta)$ بزرگتر باشد الگوریتم یادگیری به سادگی می‌تواند از نمونه‌های اضافی صرفنظر نماید و آنها را استفاده نکند. بنابراین مجموعه آموزشی بزرگتر باعث افزایش زمان یادگیری نخواهد شد و مساله یادگیری را پیچیده تر نخواهد کرد.

۲. اندازه ورودی برابر با اندازه فضای فرضیه باشد

همان گونه که پیش از این دیدیم، اندازه فضای فرضیه باعث پیچیده‌تر شدن مساله نخواهد نشد و باعث افزایش زمان اجرای الگوریتم یادگیری نخواهد شد. فضای فرضیه آستانه را می توان به‌صورت کارا یادگرفت در حالی‌که اندازه این فضا بی‌نهایت است.

۳. اندازه ورودی برابر با تعداد ابعاد فضای دامنه (تعداد ویژگی) باشد.

۴. اندازه ورودی برابر با ϵ یا δ یا تابعی از ϵ یا δ مانند $f(\epsilon, \delta)$ باشد. برای نمونه $p(\frac{1}{\epsilon}, \frac{1}{\delta})$ که $p(.)$ یک تابع چندجمله‌ای است.

۵. یک پارامتر که پیچیدگی فضای فرضیه را نشان دهد. به همین دلیل است که در تعریف یادگیری احتملا تقریبا درست از پارامتری که پیچیدگی یا قدرت توصیف فضای فرضیه را نشان بدهد استفاده می‌کنیم.

خروجی الگوریتم یادگیری یک تابع به فرم $h : \mathcal{X} \mapsto \mathcal{Y}$ است. اینکه چگونه این تابع را نمایش دهیم بسیار مهم است. برای نمونه الگوریتمی را در نظر گیرید که ورودی آن مجموعه آموزشی S است و خروجی آن توصیفی به شکل «**الگوریتم کمینه‌سازی خطا را را روی S اجرا کن**» باشد. این توصیف قابل قبول است اما همه محاسبات را به زمان آزمون منتقل می کند. برای رفع مشکل زمان اجرای یک الگوریتم را به صورت زیر تعریف می کنیم.

تعریف ۱.۵ **زمان اجرای الگوریتم یادگیری** زمان اجرای الگوریتم یادگیری عبارت است از بیشترین زمانی که برای تولید فرضیه یا تولید برچسب یک نمونه مورد نیاز است. به عبارتی دیگر، زمان اجرای یک الگوریتم یادگیری عبارت است از

$$\max\{\text{زمان مورد نیاز برای تولید برچسب نمونه } x \text{ , زمان مورد نیاز برای تولید فرضیه } h\}.$$

این تعریف به این دلیل ارایه شده است که فرضیه تولید شده به اندازه کافی کارا باشد و بتوان از فرضیه تولید شده استفاده کرد در غیر اینصورت زمان اجرای الگوریتم‌های یادگیری معتبر نخواهد بود.

حال مساله یادگیری مربع مستطیل‌ها در فضای $\mathbb{R}^n$ را در نظر گیرید. که ورودی الگوریتم کمینه‌سازی خطای تجربی یک مجموعه آموزشی با اندازه m است و خروجی فرضیه‌ای است که روش کمینه‌سازی خطای تجربی آن را تولید می‌کند.

برای نمونه می توانیم در مساله یادگیری مستطیل با اضلاع موازی محورها، حالت های زیر را در نظر بگیریم.

۱. مقدار متغیرهای ϵ یا δ را ثابت در نظر گرفته و مقدار بعدها (n) را تغییر دهیم.

۲. مقدار متغیرهای ϵ یا n را ثابت در نظر گرفته و بازه اطمینان (δ) را تغییر دهیم.

۳. مقدار متغیرهای δ یا n را ثابت در نظر گرفته و مقدار خطا (ϵ) را تغییر دهیم.

در هر سه حالت یک دنباله از مسایل خواهیم داشت. هنگامی که اندازه ورودی و پارامترهای دیگر انتخاب شدند می توانیم تحلیل مجانبی الگوریتم را انجام دهیم.

براساس مدل یادگیری احتمالا تقریبا درست، روش معقول برای تحلیل پیچیدگی محاسباتی یک الگوریتم یادگیری، استفاده از پیچیدگی نمونه‌ای آن است. به بیانی دیگر پیچیدگی نمونه‌ای را می‌توان بعنوان اندازه مفید ورودی الگوریتم دانست. همانگونه که پیش از این بیان شد پیچیدگی نمونه ای به مقدار دقت (پارامتر ϵ)، بازه اطمینان (پارامتر δ) و پیچیدگی فضای فرضیه وابسته است. بنابراین باید بگوییم که فضای فرضیه H در زمان چند جمله‌ای قابل یادگیری است اگر یک الگوریتم یادگیری A وجود داشته باشد که فضای H را با خطای ϵ و اطمینان δ یاد می‌گیرد و پیچیدگی محاسباتی آن یک تابع چندجمله‌ای براساس پیچیدگی نمونه‌ای‌اش باشد. تعریف بالا فرض می‌کند که اندازه هر نمونه ثابت است. متداول است که به صورت صریح براساس ابعاد ویژگی تحلیل پیچیدگی محاسباتی را انجام داد. در این حالت یک ضریب $\theta(n)$ که n تعداد بعدهای ورودی است در عبارت زمان اجرا ضرب می‌شود.

تعریف بالا فریبنده است بدین معنا که الگوریتم یادگیری در زمان آموزش نمونه‌ها را در حافظه ذخیره نماید و در زمانی که نمونه آزمون وارد می‌شود با روش کمینه‌سازی خطای تجربی، فرضیه را پیدا نموده و سپس برچسب نمونه آزمون را مشخص می نماید. برای جلوگیری از چنین فریبندگی، نیاز داریم که فرضیه تولید شده نیز در زمان چندجمله‌ای براساس پیچیدگی نمونه‌ای، برچسب نمونه های آزمون را تعیین نماید.

تعریف ۲.۵ **پیچیدگی محاسباتی الگوریتم‌های یادگیری** فرض کنید H فضای فرضیه روی فضای نمونه n بعدی، ϵ پارامتر دقت و δ پارامتر اطمینان باشد. همچنین فرض کنید $m_H(\epsilon, \delta)$ پیچیدگی نمونه‌ای با پارامترهای دقت ϵ و اطمینان δ باشد. آنگاه پیچیدگی محاسباتی یادگیری احتمالا تقریبا درست فضای فرضیه H از مرتبه $O(nm_H(\epsilon, \delta))$ است اگر یک الگوریتم یادگیری وجود داشته

باشد که فضای H را در زمان $O(nm_H(\epsilon,\delta))$ احتمالا تقریبا درست یاد بگیرد و زمان اجرای فرضیه تولید شده توسط این الگوریتم برای تعیین برچسب نمونه x از مرتبه $O(nm_H(\epsilon,\delta))$ باشد.

در تعریف بالا، همه پارامترهای مرتبط همانند فضای فرضیه H، ابعاد فضای نمونه n، پارامتر دقت ϵ و پارامتر اطمینان δ به کار رفته اند. درنتیجه اندازه مفید ورودی $nm_H(\epsilon,\delta)$ نیز یک عدد ثابت است. با توجه به تعریف بالا، حال یادگیری در زمان چندجمله‌ای را تعریف می نماییم. برای این تعریف ما در باره دنباله‌ای از مسایل با رشد اندازه مفید ورودی سخن می گوییم. از آنجایی اندازه مفید ورودی به فضای فرضیه H، ابعاد فضای نمونه n، پارامتر دقت ϵ و پارامتر اطمینان δ وابسته است لذا ما باید به صورت دقیق مشخص نماییم که کدام یک از اینها ثابت و کدام یک متغیر هستند. برای روشن شدن مطلب، یادگیری فضای فرضیه متناهی را در نظر بگیرید. این مساله را می توانیم با استفاده از جستجوی کورکورانه روی H با استفاده از مجموعه آموزشی با اندازه $m_H(\epsilon,\delta)$ در زمان $O(nm_H(\epsilon,\delta))$ یاد بگیریم. اگر $|H|$ را ثابت فرض کنیم و بقیه پارامترها را متغیر فرض کنیم این الگوریتم در زمان چند جمله‌ای اجرا می‌شود. اگر بقیه پارامترها را ثابت فرض کنیم و زمان اجرا را براساس $|H|$ تحلیل کنیم آنگاه زمان اجرای این الگوریتم نمایی است در حالیکه $m_H(\epsilon,\delta)$ به صورت لگاریتمی با $|H|$ رشد می‌کند. زیرا $|H|$ بر اساس n یک تابع نمایی است.

با توجه به مطالب بیان شده، یادگیری کارا به صورت زیر تعریف می‌شود.

تعریف ۳.۵ **یادگیری کارا** دنباله‌ای از مسایل یادگیری $(n_k, H_k, \epsilon_k, \delta_k)_{k=1}^{\infty}$ به صورت کارا قابل یادگیری است اگر یک چند جمله‌ای p وجود داشته باشد به گونه‌ای که برای هر k یک الگوریتم یادگیری وجود داشته باشد به گونه‌ای که فضای H_k را به صورت احتمالا تقریبا درست با پارامترهای ϵ_k و δ_k در زمان $p\left(n_k m_{H_k}\left(\epsilon_k,\delta_k\right)\right)$ یاد بگیرد.

برای نمونه در یادگیری فضای فرضیه متناهی که پیش از این بیان شد دو دنباله از مسایل یادگیری زیر را در نظر بگیرید.

۱. در دنباله نخست پارامترهای n_k، H_k و δ_k را ثابت فرض می‌کنیم که با n، H و δ نمایش می‌دهیم و مقدار $\epsilon_k = \frac{1}{k}$ در نظر می‌گیریم که متغیر است. در این حالت چندجمله‌ای به صورت $p(x) = |H|x$ و روشن است که الگوریتم کمینه سازی خطای تجربی در زمان $p\left(n_k m_{H_k}\left(\epsilon_k,\delta_k\right)\right)$ اجرا می‌شود و بنابراین این دنباله از مسایل یادگیری به صورت کارا قابل یادگیری هستند.

۲. در دنباله دوم پارامترهای n_k، ϵ_k و δ_k را ثابت فرض می‌کنیم که با n، ϵ و δ نمایش می‌دهیم و مقدار $\log|H_k| = k$ در نظر گیریم که متغیر است. در این حالت هیچ الگوریتم کمینه سازی خطای تجربی نمی‌توان پیدا کرد که در زمان $p\left(n_k m_{H_k}\left(\epsilon_k,\delta_k\right)\right)$ مساله را حل نماید و درنتیجه روش کمینه سازی خطای تجربی نمی تواند این مساله را در زمان چند جمله‌ای حل نماید.

حال با توجه به دو تعریفی که در این بخش گفته شد پیچیدگی محاسباتی الگوریتم‌های یادگیری را به صورت زیر تعریف می کنیم.

تعریف ۴.۵ **پیچیدگی محاسباتی الگوریتم‌های یادگیری** پیچیدگی محاسباتی الگوریتم‌های یادگیری در دو گام نخست مشخص می‌شوند. نخست پیچیدگی محاسباتی یک مساله یادگیری ثابت که با سه‌تایی $(\mathcal{Z}, H, \ell)$ مشخص می‌شود را محاسبه می‌کنیم که $\mathcal{Z}$ فضای نمونه، H فضای فرضیه و ℓ تابع هزینه است. سپس در گام دوم نرخ تغییرات پیچیدگی محاسباتی را با دنباله‌ای از این مسایل در نظر می‌گیریم.

۱. تابع $\mathcal{N} \mapsto (0,1)^2 : f$ ، مساله یادگیری $(\mathcal{Z}, H, \ell)$ و الگوریتم یادگیری A را در نظر بگیرید. می گوییم الگوریتم A مساله یادگیری را در زمان $O(f)$ حل می کند اگر یک ثابت c وجود داشته باشد به گونه‌ای که برای هر توزیع $\mathcal{D}$ روی $\mathcal{Z}$ و ورودی‌های $\epsilon, \delta \in (0,1)$، هنگامی که الگوریتم A مجموعه آموزشی که به صورت مستقل از توزیع $\mathcal{D}$ نمونه‌برداری شده‌اند را دریافت نماید

(آ) الگوریتم A پس از زمان حداکثر $cf(\epsilon,\delta)$ پایان می‌یابد.

(ب) خروجی الگوریتم A که با h_A نمایش داده‌می‌شود را می‌توان برای پیش‌بینی برچسب نمونه‌های جدید به‌کار برد به گونه‌ای که زمان اجرای آن حداکثر $cf(\epsilon,\delta)$ باشد.

(ج) خروجی الگوریتم A احتمالا تقریبا درست باشد یعنی با احتمال دست‌کم$\delta - ۱$ نابرابری زیر برقرار است.

$$\mathbf{R}(h_A) \leq \min_{h' \in H} \mathbf{R}(h') + \epsilon. \qquad (۲.۵)$$

۲. دنباله‌ای از مسایل یادگیری $(\mathcal{Z}_k, H_k, \ell_k)_{k=۱}^{\infty}$ را در نظر بگیرید به گونه‌ای که مساله kام با فضای دامنه $\mathcal{Z}_k$، فضای فرضیه H_k و تابع هزینه ℓ_k مشخص می گردد. فرض کنید که الگوریتم A برای حل مسایلی از این دست طراحی شده‌باشد. یک تابع $g : \mathcal{N} \times (۰,۱)^۲ \mapsto \mathcal{N}$ را در نظر بگیرید. می‌گوییم زمان اجرای الگوریتم A برای این دنباله از مسایل $O(g)$ است اگر برای همه kها، الگوریتم A مساله $(\mathcal{Z}_k, H_k, \ell_k)$ را در زمان $O(f_k)$ حل می‌کند به گونه‌ای که $f_k : (۰,۱)^۲ \mapsto \mathcal{N}$ به صورت $f_k(\epsilon,\delta) = g(n,\epsilon,\delta)$ تعریف می‌شود.

می گوییم که الگوریتم A نسبت به دنباله $(\mathcal{Z}_k, H_k, \ell_k)$ یک الگوریتم کارا است اگر زمان اجرای آن از مرتبه $O(p(n,۱/\epsilon,۱/\delta))$ باشد که p یک چندجمله‌ای براساس پارامترهایش است.

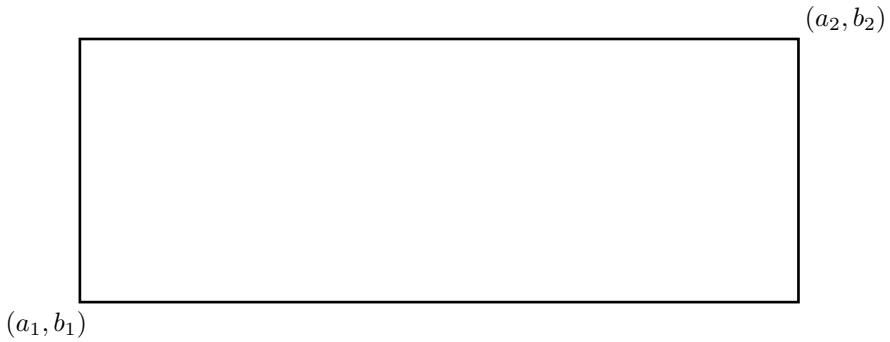
از تعریف بالا روشن است که کارایی الگوریتم به شیوه شکستن مساله به دنباله‌ای از مسایل وابسته است. برای نمونه، یادگیری یک فضای فرضیه متناهی را در نظر بگیرید. همان گونه که پیش از این نشان دادیم روش کمینه سازی خطای تجربی روی فضای فرضیه H یادگیری را تضمین می کند اگر اندازه مجموعه آموزشی دست کم $m_H(\epsilon,\delta) = \log(|H|/\delta)/\epsilon^۲$ باشد. فرض کنید که زمان ارزیابی هر فرضیه به زمان ثابتی نیاز داشته باشد، الگوریتم جستجوی کورکورانه در زمان $O(|H|m_H(\epsilon,\delta))$ مساله را یاد می‌گیرد. برای هر فضای فرضیه متناهی اگر دنباله مسایل را به صورت $|H_k| = k$ تعریف کنیم الگوریتم A کارا است. اما اگر دنباله مسایل را به صورت $|H_k| = ۲^k$ تعریف کنیم آنگاه پیچیدگی نمونه‌ای هنوز به یک تابع چندجمله‌ای براساس پارامترهایش است اما پیچیدگی محاسباتی الگوریتم A به صورت نمایی رشد با k رشد می کند بنابراین دیگر الگوریتمی کارا نیست.

۲.۵ پیاده سازی قاعده کمینه سازی خطای تجربی

برای هر فضای فرضیه Hی، قاعده کمینه سازی خطای تجربی یک رویکرد متداول یادگیری است. برای نمونه، پیش از این دیدیم که اگر برای یک مساله دسته بندی دودویی، یادگیری امکان‌پذیر باشد با قاعده کمینه سازی خطای تجربی نیز امکان‌پذیر است. در ادامه این بخش، پیچیدگی محاسباتی پیاده سازی قاعده کمینه سازی خطای تجربی برای فضاهای فرضیه مختلف را بررسی می کنیم. برای فضای دامنه $\mathcal{Z}$، فضای فرضیه H و تابع هزینه ℓ قاعده کمینه سازی خطای تجربی به صورت زیر تعریف می‌شود. برای هر مجموعه آموزشی متناهی $S \sim \mathcal{D}^m$ ، فرضیه‌ای را تولید می‌کند که خطای تجربی $\frac{۱}{|S|}\sum_{z \in S} \ell(h,z)$ را کمینه نماید.

۱.۲.۵ فضای فرضیه متناهی

محدود کردن فضای فرضیه به یک فضای متناهی، بسیار منطقی است زیرا هنگامی که می خواهیم یک فرضیه را در کامپیوتر نشان دهیم باید گسسته سازی صورت پذیرد. برای نمونه فضای خطوط $ax+b$ را درنظر بگیرید. برای نشان دادن پارامترهای این خط باید آن‌ها را در حافظه کامپیوتر ذخیره‌سازی نماییم برای نمونه هرکدام از پارامترهای a و b را با ۱۶ بیت نمایش دهیم در این صورت تعداد خطوطی که با این حافظه



شکل ۲.۵ نمایش یک مستطیل در فضای دوبعدی

می‌توان نمایش داد متناهی خواهد بود. نمونه دیگر فضای فرضیه‌ای را در نظر بگیرید که بتوان با ۱۰۰۰ بایت برنامه $C++$ آن‌ها را نوشت. همان گونه که در بخش های پیشین نشان دادیم، کران بالای پیچیدگی نمونه‌ای یادگیری فضای فرضیه متناهی برابر است با $m_H(\epsilon, \delta) = \log(c|H|/\delta)/\epsilon^c$ که برای حالت تحقق‌پذیر $c = 1$ و برای حالت تحقق‌ناپذیر $c = 2$ است. برای نمونه برای مثال پیاده‌سازی با زبان $C++$، تعداد فرضیه ها برابر است با 2^{1000} در حالی‌که پیچیدگی نمونه ای برابر است با $m_H(\epsilon, \delta) = c(1000 + \log \frac{c}{\delta})/\epsilon^c$. یک رویکرد سراست برای پیاده سازی قاعده کمینه سازی خطای تجربی برای فضای فرضیه متناهی جستجوی کورکورانه است. اگر فرض کنید که محاسبه مقدار $\ell(h, z)$ برای یک نمونه در زمان ثابت k انجام پذیرد. زمان اجرای الگوریتم کمینه سازی قاعده تجربی براب است با $k|H|m$ که m اندازه مجموعه آموزشی است. اگر مقدار m برابر کران بالای پیچیدگی محاسباتی باشد در این صورت زمان اجرای این الگوریتم برابر است با $ck|H|\log(c|H|/\delta)/\epsilon^c$. این رابطه نشان می دهد که زمان اجرای الگوریتم به صورت خطی به اندازه فضای فرضیه $(|H|)$ وابسته است و این وابستگی خطی سبب می شود که این الگوریتم برای فضاهای فرضیه بزرگ مناسب نباشد.اگر دنباله مسایل $(\mathcal{Z}_k, H_k, \ell_k)_{k=1}^{\infty}$ را به گونه‌ای تعریف کنیم که $\log(|H_k|) = k$، زمان اجرای جستجوی کورکورانه از مرتبه نمایی است. به همین جهت معمولا از فضاهای فرضیه دیگر مانند فضای ابرصفحه‌ها استفاده می‌شود. نکته ای که بسیار مهم است این است که کارایی پایین یک روش پیاده‌سازی قاعده کمینه سازی خطای تجربی به این معنا نیست که نمی‌توان این قاعده را به صورت کارا پیاده سازی نمود. در ادامه این بخش مثال های دیگری را بررسی می کنیم که می‌توان این قاعده را به صورت کارا پیاده‌سازی نمود.

۲.۲.۵ مستطیل هایی با اضلاع موازی محورهای مختصات

پیش از این درباره فضای مستطیل‌های با اضلاع موازی محورهای مختصات سخن به میان آمد. با توجه به شکل زیر می‌توانیم یک مستطیل در فضای n بعدی را به $2n$ پارامتر نشان داد.

بنابراین H_n را که عبارت است از فضای همه مستطیل‌های با اضلاع موازی محورهای مختصات در فضای n بعدی را به صورت زیر در تعریف می‌کنیم.

$$H_n = \left\{h_{(a_1,a_2,\ldots,a_n,b_1,b_2,\ldots,b_n)} : \forall i, a_i \leq b_i\right\} \tag{۳.۵}$$

که

$$h_{(a_1,a_2,\ldots,a_n,b_1,b_2,\ldots,b_n)}(x) = \begin{cases} 1 & \text{if } \forall i, x_i \in [a_i, b_i] \\ 0 & \text{otherwise.} \end{cases} \tag{۴.۵}$$

نخست حالت سازگار را در نظر بگیرید یعنی در فضای فرضیه یک فرضیه $h \in H_n$ وجود دارد به‌گونه‌ای که برای همه نمونه‌های آموزشی $(x,y) \in S$ شرط $h(x) = y$ برقرار است. حال پیاده‌سازی قاعده کمینه سازی خطای تجربی برای این فضا را به این صورت در نظر بگیرید که a_i برابر باشد با کوچکترین مقدار بعد iام نمونه های مثبت و b_i برابر باشد با بزرگترین مقدار بعد iام نمونه های مثبت. به بیانی دیگر برای همه iها داشته باشیم

$$a_i = \min\{x_i : (x_i, ۱) \in S\}$$
$$b_i = \max\{x_i : (x_i, ۱) \in S\}$$

به سادگی می‌توان نشان داد که خطای تجربی برابر صفر است و فرضیه تولید شده با داده‌های آموزشی سازگار است. برای پیدا کردن هرکدام از a_iها و b_iها نیاز به زمان $O(m)$ است و چون n بعد داریم لذا زمان اجرای الگوریتم $O(nm)$ است. حال حالتی را در نظر بگیرید که فرضیه سازگار وجود نداشته باشد و به دنبال فرضیه‌ای هستیم که خطای آموزش را کمینه نماید. نشان داده شده است که کمینه‌سازی خطای تجربی در این حالت یک مساله NP-Hard است و بنابراین راه حل چندجمله ای ندارد. اما اگر فضای فرضیه را محدود نماییم برای نمونه بعد n را ثابت درنظر گیریم الگوریتم کارا برای حل مساله وجود دارد. یعنی الگوریتم یادگیری احتمالا تقریبا درست بدون پیش فرض وجود دارد که در زمان چند جمله‌ای با $\frac{۱}{\epsilon}$ و $\frac{۱}{\delta}$ اجرا می شود اما تابعی چندجمله‌ای از n نیست. برای نمونه برای پیدا کردن مستطیل در n بعد حداکثر به $۲n$ نمونه است بنابراین می‌توانیم همه زیر مجموعه‌های $۲n$ عضوی از مجموعه m عضوی آموزش را در نظر بگیریم و با استفاده از جستجوی کورکورانه مساله را حل نماییم. بدیهی است که الگوریتم فرضیه با کمینه خطای تجربی را پیدا کی کند. زمان اجرای این الگوریتم برابر است با $m^{O(n)}$. در نتیجه این الگوریتم در زمان چندجمله ای نسبت به m و در زمان نمایی نسبت به n اجرا می شود.

۳.۲.۵ ترکیب عطفی تعدادی ویژگی دودویی

یک ترکیب عطفی تعدادی ویژگی دودویی نگاشتی از $X = \{۰,۱\}^n$ به $Y = \{۰,۱\}$ است و می توان به صورت گزاره $x_{i_۱} \wedge x_{i_۲} \wedge \ldots x_{i_k} \wedge \bar{x}_{j_۱} \wedge \bar{x}_{j_۲} \ldots \wedge \bar{x}_{j_r}$ نشان داد که $i_۱, i_۲, \ldots, i_k, j_۱, j_۲ \ldots, j_r \in \{۱, ۲, \ldots, n\}$. تابعی که این گزاره نشان می دهد برابر است با

$$h(x) = \begin{cases} ۱ & \text{if } x_{i_۱} = \ldots = x_{i_k} = ۱ \text{ and } = x_{j_۱} = \ldots = x_{j_r} = ۰ \\ ۰ & \text{otherwise.} \end{cases} \quad (۵.۵)$$

حال مجموعه چنین توابعی را با H_n نشان می دهیم. نخست پیاده سازی قاعده کمینه سازی خطا را در حالتی که فرضیه سازگار وجود دارد در نظر بگیرید. نشان می دهیم که برای این حالت الگوریتمی کارا وجود دارد. این الگوریتم نخست فرضیه $h = (x_۱ \wedge \bar{x}_۱) \wedge (x_۲ \wedge \bar{x}_۲) \wedge \ldots \wedge (x_n \wedge \bar{x}_n))$ در نظر می‌گیرید. بدیهی است که این فرضیه به همه نمونه هایی که دریافت می کند برچسب صفر انتساب می دهد. این الگوریتم با دریافت نمونه های مثبت، حروفی(ویژگی ها یا نقیض آنها) که با نمونه سازگار نباشند را از فرضیه حذف می کند و با دریافت نمونه های منفی کاری انجام نمی‌دهد. زمان اجرای این الگوریتم $O(nm)$ است. اما برای حالتی که فرضیه سازگار موجود نباشد الگوریمی با زمان اجرای چندجمله‌ای براساس n و m وجود ندارد.

۴.۲.۵ یادگیری کلاس مفاهیم 3-term DNF

در این مثال نشان می‌دهیم که تعمیم ترکیب عطفی ویژگی های دودویی منجر به بغرنج شدن حل مساله کمینه‌سازی خطای تجربی می‌شود. فضای فرضیه های H_n شامل همه 3-term DNFها (یا ترکیب سه عبارت با طول دلخواه) باشد. برای نمونه تابع $h(x_۱, \ldots, x_n) =$

$(x_۲ \wedge x_۷ \wedge x_{۱۰}) \vee (x_۱ \wedge x_۵) \vee (x_۲ \wedge x_۷)$ یک 3-term DNF است. تعداد توابع 3-term DNF برابر است با $۳^{۳^n}$ و بنابراین اندازه H_n برابر است با $۳^{۳^n}$ (یا $\log|H_k| = \theta(n)$) و در نتیجه کران بالای پیچیدگی نمونه‌ای آن برابر است با $۳n\log(۳/\delta)/\epsilon$.

در این حالت $h(x)$ را می توانیم به صورت $h(x) = A_۱(x) \vee A_۲(x) \vee A_۳(x)$ نمایش دهیم و زمانی $h(x) = ۱$ است که دست کم یکی از عبارت های $A_۱(x)$ یا $A_۲(x)$ و یا $A_۳(x)$ صفر نباشند. نشان داده شده است که هیچ الگوریتم چندجمله برای حل این مساله وجود ندارد.

۳.۵ یادگیری مستقل از نمایش، سخت نیست

در بخش پیش نشان دادیم که هیچ الگوریتمی با پیچیدگی زمانی چند جمله برای حل مساله 3-term DNF وجود ندارد. در این بخش نشان می دهیم که این فضا را با کمک الگوریتم کمینه سازی خطا با استفاده از فضای بزرگتری می توانیم یاد بگیریم. این نتیجه تناقضی با نتیجه بخش پیشین ندارد. زیرا الگوریتم یادگیری فرضیه ای را تولید می‌کند که به فضای فرضیه تعریف شده اصلی تعلق ندارد. در اینجا فضای توابع 3-term DNF را با فضای بزرگتری که با کارایی بیشتری قابل یادگیری است جایگزین می کنیم. یعنی فرضیه‌ای که الگوریتم یادگیری تولید می نماید الزاما یک 3-term DNF نیست. به همین دلیل نام یادگیری مستقل از نمایش یا یادگیری نامناسب [۱] بر روی آن گذاشته شده است. برای حل مساله از این واقعیت استفاده می‌کنیم که هر 3-term DNF یک 3-term DNF معادل دارد. برای پیدا کردن این معادل، $\vee$ را روی $\wedge$ توزیع می‌کنیم. با توزیع $\vee$ روی $\wedge$، فضای ویژگی از n به $(۲n)^۳$ تغییر می کند. با توجه به اینکه فرض کردیم فرضیه سازگار وجود دارد لذا با الگوریتم کمینه سازی خطای تجربی می توانیم این فرضیه را پیدا کنیم . کران بالای پیچیدگی نمونه ای آن برابر است با $n^۳\log(۱/\delta)/\epsilon$ و در نتیجه زمان اجرا چندجمله ای براساس n است. این مثال نشان می دهد که می توانیم با تغییر نمایش مساله راحت‌تر یادگرفته شود.

۴.۵ سختی یادگیری

پیش از این در باره پیچیدگی محاسباتی یادگیری سخن به میان آمد. همچنین نشان دادیم که سختی پیاده سازی الگوریتم کمینه سازی خطای تجربی برای فضای H الزاما به این معنا نیست که فضای فرضیه H قابل یادگیری نیست. حال پرسش این است که چگونه می توانیم نشان دهیم که یک مساله یادگیری از نظر محاسباتی یک مساله سخت است. یک رویکرد استفاده از فرضیات مورد استفاده در رمزنگاری است. از یک رویکرد، می توان یادگیری را مخالف رمزنگاری دانست. در یادگیری ما فرض می کنیم اگر به مجموعه آموزشی که توسط یک مفهوم برچسب خورده است دسترسی داشته باشیم می توانیم با خطای قابل قبولی مفهوم را بیادبگیریم و مفهوم را استخراج نماییم در حالیکه در رمز نگاری هدف این است که یک فرد نمی تواند با دریافت تعدادی پیام، محتوی آن ها را استخراج نماید. بنابراین می توانیم نتایج موجود در رمزنگاری را به حوزه یادگیری منتقل نماییم. بدبختانه در حال حاضر کسی نمی تواند ادعا نماید که یک پروتکل رمزنگاری امن است.

در ادامه این بخش ما از نتایج بدست آمده در رمزنگاری برای نشان دادن یادگیری استفاده می کنیم. در رمز نگاری برای اثبات اینکه یک پروتکل رمزنگاری امن است از فرضیات رمز نگاری استفاده می شود. بسیاری از روش های رمز نگاری فرض می کنند که یک تابع یک طرفه $f : \{۰,۱\}^n \mapsto \{۰,۱\}^n$ وجود دارد که به سادگی محاسبه می‌شود. اما محاسبه وارون آن سخت است. بطور رسمی تابع f را می توان در زمان چند جمله‌ای براساس n محاسبه نمود اما برای هر الگوریتم تصادفی با زمان اجرای چندجمله‌ای A و برای هر چندجمله‌ای $p(.)$

[1] Improper learning

نابرابری زیر برقرار است.

$$\mathbb{P}\left[f(A(f(x))) = f(x)\right] < \frac{۱}{p(n)}, \tag{۶.۵}$$

که احتمال براساس انتخاب تصادفی x براساس توزیع یکنواخت روی $\{۰,۱\}^n$ و تصادفی بودن الگوریتم A محاسبه می‌شود.

برای آینکه محاسبه وارون آن سخت باشد از یک کلید استفاده می‌شود و شخصی که کلید s_n را داشته باشد به سادگی می تواند تابع f را محاسبه و شخصی که کلید s_n را را نداشته باشد به سختی می‌تواند وارون تابع f را محاسبه نماید.

فرض کنید که F_n خانواده ای از توابع رمزنگاری روی $\{۰,۱\}^n$ باشد که در زمان چند جمله ای قابل محاسبه هستند. اگر ما الگوریتم رمزنگاری را ثابت فرض کنیم و یک کلید s_n که یک تابع درون F_n را نشان می‌دهد و رشته x به آن بدهیم در زمان چندجمله‌ای می تواند $f(x)$ را محاسبه نماید. حال مساله یادگیری وارون این فضا $H_n^F = \{f^{-۱} : f \in F_n\}$ را در نظر بگیرید. از آنجایی که وارون هر تابع در این فضا را می توان بوسیله کلید s_n محاسبه نمود. اندازه این کلید یک چندجمله ای براساس n (که با $p(n)$ نمایش داده می‌شود) است. بنابراین فضای H_n^F را می توان براساس کلید پارامتری نموده و اندازه آن حداکثر $۲^{p(n)}$ است و در نتیجه پیچیدگی نمونه‌ای آن چندجمله‌ای است.

می‌توان نشان داد که هیچ یادگیر کارایی برای حل این مساله وجود ندارد. اگر چنین یادگیری وجود داشت می توانستیم از روی رشته های فضای $\{۰,۱\}^n$ با توزیع یکنواخت نمونه برداری نمود و تابع $f(x)$ را برای هر نمونه x محاسبه نمود و سپس یک مجموعه آموزشی با اندازه چندجمله‌ای از زوج‌های $(f(x), x)$ تولید نمود و در پایان با تقریب تابع $f^{-۱}$ را بدست آورد که یکطرفه بودن f را نقض می کند.

واژه‌نامه فارسی به انگلیسی

ب

بیش برازش ..... Overfitting

پ

پیچیدگی نمونه‌ای ..... Sample complexity

ت

توابع عطفی ..... Conjunctions
- عطف یکنوا ..... Monotone conjunctions
- فرم نرمال عطفی درجه k ..... k-CNF

توابع فصلی ..... Disjunctions
- فرم نرمال فصلی دو لفظی ..... 2-Term DNF

خ

خصوصیات ..... Attributes

خطا ..... Error
- خطای برآورد ..... Estimation error
- خطای تجربی ..... Empirical Error
- خطای تقریب ..... Approximation Error
- خطای واقعی ..... True Error

د

دانش پیشین ..... Prior knowldege

دسته بند ..... Classifier

دسته‌ای ..... Batch

ف

فرضیه ..... Hypothesis

فضای نمونه ..... Domain set

ک

کران اجتماعات ..... Union bound

م

مجموعه آموزشی ..... Training set

مجموعه برچسب ها ..... Label set

مدل احتمالاً تقریبا درست ..... PAC model

مفهوم ..... Concept

اندازه مفهوم ..... Concept size

کلاس مفاهیم ..... Concept class

ن

نمونه ..... Example

نمونه های آزمایشی ..... Test examples

و

ویژگی ها ..... Features

ویژگی همگرایی نایکنواخت ..... Nonuniform convergence property

ی

یادگیر عمومی ..... Universal learner

یادگیری نایکنواخت ..... Nonuniform learnability

نمایه