

فصل پنجم: ارزیابی فرضیه‌ها

ارزیابی تجربی دقت فرضیه‌ها اساس یادگیری ماشین است. این فصل معرفی‌ای بر متدهای تخمین دقت فرضیه‌ها با تاکید بر سه سؤال زیر ارائه می‌کند. اول اینکه با در دست داشتن دقت فرضیه بر روی مجموعه‌ی محدودی از داده‌ها، این تقریب چقدر به دقت بر روی نمونه‌های جدید نزدیک است؟ دوم اینکه با دانستن اینکه یک فرضیه از فرضیه‌های دیگر دقت بیشتری بر روی تعدادی نمونه دارد، چقدر احتمال دارد که این فرضیه در کل از فرضیه‌های دیگر دقت بیشتری داشته باشد؟ سوم اینکه زمانی که داده‌ها محدود است بهترین روش برای استفاده‌ی این داده‌ها هم برای آموزش و هم برای ارزیابی چیست؟ زیرا که تعداد محدود نمونه‌ها ممکن است معرف توزیع کلی نمونه‌ها نباشد و در بدست آوردن دقت فرضیه بر روی تمامی نمونه‌ها گمراه کننده باشند. متدهای آماری، با فرض‌هایی که درباره‌ی توزیع داده‌ها انجام می‌دهند، اجازه می‌دهند تا حداکثر اختلاف بین دقت مشاهده شده بر روی داده‌های موجود و دقت واقعی روی کل توزیع داده‌ها را محاسبه کنیم.

5.1 انگیزه

در بسیاری از موارد ارزیابی فرضیه‌ی یاد گرفته شده با حداکثر دقت ممکن بسیار مهم است. یکی از دلایل این اهمیت، تشخیص قابل استفاده بودن فرضیه است (مشخص شدن این است که آیا این فرضیه را به کار ببریم یا خیر). برای مثال، زمانی که از یک پایگاه داده‌ی محدود برای بررسی تأثیر داروها استفاده می‌کنیم، دانستن دقت فرضیه‌ی یاد گرفته شده بسیار مهم است. دلیل دوم اهمیت ارزیابی فرضیه‌ها این است که ارزیابی فرضیه‌ها عنصر داخلی بسیاری از الگوریتم‌های یادگیری است. برای مثال، در هرس کردن درخت‌های تصمیم برای حل کردن مشکل **overfit** باید تأثیر هرس بر دقت درخت حاصل را در هر مرحله بدانیم. بنابراین درک وجود خطای ذاتی در تخمین دقت درخت‌ها قبل از هرس بعد از هرس اهمیت بسیار دارد.

تخمین دقت یک فرضیه هنگامی که تعداد داده‌ها زیاد است بسیار ساده خواهد بود. با این حال، زمانی که لازم است با تعداد محدودی داده فرضیه‌ای را یاد بگیریم و ارزیابی کنیم دو مشکل اساسی پیش می‌آید:

- بایاس در تخمین. اول اینکه دقت فرضیه بر روی نمونه های آموزشی اغلب معیار ضعیفی برای دقت فرضیه بر داده های جدید است. زیرا که فرضیه از همین داده ها به وجود آمده است، پس این نمونه های تخمینی بایاس دار و خوش بینانه از دقت فرضیه بر داده های جدید می زنند. این مشکل بیشتر مواقعی پیش می آید که فضای فرضیه ای، فضایی کامل است و به فرضیه اجازه می دهد بر روی نمونه های آموزشی $overfit$ شود. برای بدست آوردن تخمینی بدون بایاس از دقت فرضیه بر روی داده های جدید، معمولاً فرضیه را بر روی دسته نمونه های مجزایی از نمونه های آموزشی (دسته ی تست) می سنجیم.
- اختلاف در تخمین. دوم اینکه اگر دقت فرضیه را بر روی دسته ی تست که بایاس ندارند بسنجیم، با این حال امکان دارد که این دقت بدست آمده با دقت واقعی اختلاف داشته باشد، این اختلاف به چگونگی انتخاب دسته ی تست وابسته است. با کاهش اندازه ی دسته ی تست این میزان خطای احتمالی نیز افزایش می یابد.

در این فصل به متد های ارزیابی فرضیه های یاد گرفته شده، متدهای مقایسه ی دقت دو فرضیه، و متد های مقایسه ی دقت دو الگوریتم یادگیری مختلف هنگامی با داده ها محدودند می پردازیم. اکثر این مباحث بر پایه ی قوانین پایه ای آماری و تئوری نمونه برداری¹ هستند، البته در طول این فصل فرض شده که خواننده هیچ اطلاعات قبلی ای در مورد مباحث پیچیده ی آماری ندارد. تحقیق بر روی تست های آماری بررسی فرضیه ها خیلی وسیع است. این فصل خلاصه ای مقدمه ای از این تحقیقات آماری با تاکید بر قسمت هایی که بیشترین رابطه را با یادگیری و تخمین و مقایسه ی دقت فرضیه ها دارد را نیز ارائه می کند.

5.2 تخمین دقت فرضیه ها

زمانی که دقت یک فرضیه را تخمین می زنیم هدف دقت فرضیه برای دسته بندی نمونه های جدید است، علاوه بر این علاقه داریم که احتمال خطا در تخمین این دقت را نیز بسنجیم (چه میزان خطایی را باید در این تخمین در نظر گرفت).

در تمام طول این فصل از نماد گذاری ذیل برای مسایل یادگیری استفاده خواهیم کرد. فضای نمونه های X (برای مثال مجموعه ی کل افراد جامعه) وجود دارد که تابع هدف های متفاوتی (مثل افرادی که می خواهند امسال تخته اسکی جدید بخرند) روی آن ها تعریف می شوند. ما فرض می کنیم که تعداد تکرار اعضای مختلف X مساوی نیست. راه حل ساده برای در نظر گرفتن این فرض این است که توزیع احتمال مجهول D را که بر روی X تعریف شده برای تعداد تکرار نمونه ها در نظر بگیریم (این تابع توزیع ممکن است برای افراد 19 ساله خیلی بیشتر از افراد 109 ساله باشد). توجه داشته باشید که D هیچ اطلاعاتی در مورد مثبت یا منفی بودن نمونه x به ما نمی دهد؛ این توزیع فقط احتمال برخورد با نمونه x را به ما می دهد. کار یادگیری، یادگیری تابع هدف f با استفاده از فضای فرضیه های ممکن H است. نمونه های آموزشی تابع هدف f توسط یک معلم به یادگیر داده می شود. معلم جداگانه بر اساس توزیع D نمونه ها را انتخاب می کند سپس نمونه ی x را با مقدار تابع هدف $f(x)$ به یادگیر می دهد.

برای تصور، تابع هدف "افرادی که می خواهند امسال تخته اسکی جدید بخرند" را با نمونه های آموزشی ای که از تحقیقی که از افرادی که به پیست اسکی وارد می شوند بدست آمده در نظر بگیرید. در این مثال، فضای نمونه های X کل افراد جامعه است، این نمونه ها با ویژگی هایی نظیر سن، شغل، تعداد دفعات اسکی در سال و غیره توصیف می شوند. توزیع D برای هر فرد x احتمال اینکه فرد بعدی ای باشد که وارد پیست می شود را می دهد. تابع هدف $f: X \rightarrow \{0,1\}$ افراد را بر اساس اینکه قصد خرید تخته اسکی جدید دارند دسته بندی می کند.

¹ sampling theory

با این نماد گذاری کلی ما به دنبال جواب دو سؤال زیر هستیم:

1. برای فرضیه‌ی h و یک مجموعه نمونه n عضوی که اعضایش با توزیع احتمال \mathcal{D} انتخاب شده‌اند، بهترین تخمین دقت h بر روی نمونه‌های جدیدی که با همان توزیع انتخاب می‌شوند چقدر است؟
2. خطای احتمالی این تقریب دقت چقدر است؟

5.2.1 خطای نمونه‌ای و خطای واقعی

برای جواب به سؤال‌های مطرح شده، لازم است که ابتدا تفاوت بین دو دقت یا خطا را بیان کنیم. یکی نسبت خطای فرضیه بر روی نمونه‌های موجود است. دیگری نسبت خطای فرضیه بر روی کل توزیع نامعلوم \mathcal{D} از نمونه‌هاست. این دو خطا را به ترتیب خطای نمونه‌ای¹ و خطای واقعی² می‌نامیم.

خطای نمونه‌ای یک فرضیه بر اساس مجموعه نمونه‌های S از X نسبتی از S است که فرضیه اشتباه دسته بندی می‌کند:

تعریف: خطای نمونه‌ای ($error_S(h)$) برای فرضیه‌ی h بر اساس تابع هدف f و مجموعه نمونه‌های S به شکل زیر تعریف می‌شود:

$$error_S(h) \equiv \frac{1}{n} \sum_{x \in S} \delta(f(x), h(x))$$

در این رابطه n تعداد نمونه‌های S و $\delta(f(x), h(x))$ اگر $f(x) \neq h(x)$ یک است و در غیر این صورت صفر می‌باشد.

خطای واقعی یک فرضیه احتمال این است که فرضیه نمونه‌ای که با توزیع \mathcal{D} انتخاب شده را اشتباه دسته بندی می‌کند.

تعریف: خطای واقعی ($error_{\mathcal{D}}(h)$) برای فرضیه‌ی h و بر اساس تابع هدف f و توزیع \mathcal{D} احتمال این است که h نمونه‌ای انتخابی با توزیع \mathcal{D} را اشتباه دسته بندی کند.

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

در این رابطه $\Pr_{x \in \mathcal{D}}$ احتمال را برای x که با توزیع \mathcal{D} انتخاب شده باشد نشان می‌دهد.

همیشه ما قصد داریم که $error_{\mathcal{D}}(h)$ را برای فرضیه پیدا کنیم، زیرا که این میزان خطایی که در دسته بندی نمونه‌های جدید وجود دارد را بیان می‌کند. با این وجود، ما فقط می‌توانیم مقدار $error_S(h)$ را با داشتن مجموعه‌ی S اندازه گیری کنیم. سؤال اصلی این است که " $error_S(h)$ چقدر برای تخمین $error_{\mathcal{D}}(h)$ مناسب است؟"

¹ sample error

² true error

5.2.2 بازه های اطمینان برای فرضیه های گسسته مقدار

چگونه به سؤال " $error_S(h)$ با چه میزان دقت $error_D(h)$ را تخمین می‌زند؟" در زمانی که h تابعی گسسته مقدار است پاسخ می‌دهیم؟ به عبارت دقیق‌تر، فرض کنید که می‌خواهیم خطای واقعی را برای فرضیه‌ی گسسته مقدار h را با استفاده از مجموعه نمونه های S در شرایط زیر تعیین کنیم:

- مجموعه‌ی S شامل n نمونه ای است که مستقل از یکدیگر و مستقل از h هستند که با توجه به \mathcal{D} انتخاب شده‌اند.
- $N \geq 30$
- فرضیه‌ی h ، r تعداد نمونه را اشتباه دسته بندی می‌کند ($error_S(h) = r/n$)

در چنین شرایطی، تئوری آمار به ما اجازه می‌دهد تا نتایج زیر را بگیریم:

1. بدون داشتن اطلاعات بیشتر، محتمل‌ترین مقدار $error_D(h)$ ، $error_S(h)$ است.

2. با احتمال تقریباً 95٪ خطای واقعی $error_D(h)$ در بازه‌ی زیر است:

$$error_S(h) \pm 1.96 \sqrt{\frac{error_S(h)(1 - error_S(h))}{n}}$$

برای تصور، مجموعه‌ی داده های S را با $n = 40$ نمونه و فرضیه‌ی h را با $r = 12$ خطا بر روی این نمونه‌ها در نظر بگیرید. در این مثال، خطای نمونه ای $error_S(h) = 12/40 = 0.3$ است. بدون داشتن اطلاعات بیشتر، بهترین تخمین برای $error_D(h)$ همان مقدار 0.3 است. با این وجود، انتظار نمی‌رود که این تخمین، تخمین کاملی از خطای واقعی باشد. اگر دسته‌ی دیگری از نمونه‌ها مثل S' که 40 نمونه دارد داشته باشیم، قاعدتاً انتظار داریم که $error_{S'}(h)$ تقریباً با $error_S(h)$ مساوی باشد. اختلاف احتمالی این دو مقدار به چپش دو مجموعه‌ی S و S' وابسته است. در واقع اگر این آزمایش را بارها تکرار کنیم و در هر بار تکرار از مجموعه‌ی 40 نمونه ای S_i استفاده کنیم، به این نتیجه خواهیم رسید که تقریباً در 95٪ این آزمایشات بازه‌ی محاسبه شده خطای واقعی را شامل می‌شود. به همین دلیل به این بازه، بازه‌ی اطمینان 95 درصدی تخمین خطای واقعی $error_D(h)$ می‌گویند. در مثال فعلی که در آن $r = 12$ و $n = 40$ است، بازه‌ی 95٪ بر اساس داده های بالا $0.30 \pm (1.96 \cdot 0.07) = 0.30 \pm 0.14$ خواهد بود.

رابطه ای که در بالا برای بازه‌ی اطمینان 95 درصدی بیان شد را می‌توان برای اطمینان N درصدی تعمیم داد. مقدار ثابت 1.96 که در تعریف رابطه آمده برای اطمینان 95 درصدی است، با تغییر دادن این ثابت، z_N می‌توان بازه‌ی اطمینان $N\%$ را محاسبه کرد. رابطه‌ی کلی محاسبه‌ی بازه‌ی اطمینان $N\%$ برای خطای $error_D(h)$ در زیر آمده است:

$$error_S(h) \pm z_N \sqrt{\frac{error_S(h)(1 - error_S(h))}{n}} \quad (5.1)$$

در این رابطه ثابت z_N بر حسب درصد اطمینان تغییر می‌کند. مقادیر نظیر بعضی درصدها در جدول 5.1 آمده است.

خطای N	50%	68%	80%	90%	95%	98%	99%
درصدی:							

ثابت z_N : 0.67 1.00 1.28 1.64 1.96 2.33 2.58

جدول 5.1 مقادیر z_N برای بازه‌ی دوطرفه‌ی اطمینان $N\%$.

بنابراین همان طور که به راحتی محاسبه شد، بازه‌ی اطمینان $error_D(h)$ (برای $r = 12$ و $n = 40$) ± 0.30 (1.96-0.07) است، به راحتی می‌توان بازه‌ی اطمینان 68% را نیز محاسبه کرد، این بازه ± 0.30 (1.0-0.07) است. توجه دارید که منطقی است که بازه‌ی 68% از بازه‌ی 95% کوچک‌تر باشد زیرا که احتمال وجود $error_D(h)$ با کاهش طول بازه کاهش می‌یابد.

رابطه‌ی 5.1 نحوه‌ی محاسبه‌ی بازه‌ی اطمینان یا ستون‌های خطا¹ را برای تخمین $error_D(h)$ بر اساس مقدار $error_S(h)$ نشان می‌دهد. در استفاده از این رابطه همیشه باید در نظر داشت که این مقادیر فقط برای فرضیه‌های گسسته مقدار مطرح می‌شود و مجموعه‌ی نمونه‌ی S به صورت تصادفی با همان توزیع احتمالی که نمونه‌های جدید با آن انتخاب می‌شوند انتخاب شده و همچنین فرض می‌شود که داده‌ها از فرضیه‌ای که با آن تست می‌شوند مستقلند. همچنین باید در نظر داشت که این رابطه فقط تخمین بازه‌ی اطمینان است، با این وجود این تخمین زمانی که تعداد نمونه‌های S بیش از 30 است و $error_S(h)$ خیلی نزدیک 0 یا 1 نیست دقت خوبی دارد. به عبارت دقیق‌تر این رابطه زمانی که نامساوی زیر صادق است دقت خوبی دارد:

$$n \cdot error_S(h) \cdot (1 - error_S(h)) \geq 5$$

در بالا خلاصه‌ی فرایند محاسبه‌ی بازه‌ی اطمینان برای فرضیه‌های گسسته مقدار آورده شده است. قسمت بعدی توجیه آماری این فرایند را بیان می‌کند.

5.3 اساس تئوری نمونه برداری

در این قسمت ایده‌های اصلی آماری و تئوری نمونه برداری را معرفی خواهیم کرد، این ایده‌ها شامل توزیع‌های احتمال، امید ریاضی، واریانس، توزیع‌های دو جمله‌ای و نرمال و بازه‌های یک طرفه و دو طرفه می‌شود. آشنایی ابتدایی با این مفاهیم برای درک ارزیابی فرضیه‌ها و الگوریتم‌های یادگیری اساسی است. از آن مهم‌تر اینکه این مفاهیم محیطی برای درک مشکلات یادگیری ماشین مثل overfit و رابطه‌ی بین تعمیم موفق و تعداد نمونه‌ها آموزشی فراهم می‌کنند. خواندگانی که از قبل با این مفاهیم آشنایی دارند می‌توانند بدون لطمه وارد شدن به همبستگی کتاب بدون خواندن این قسمت رد شوند. مفاهیم کلیدی‌ای که در این بخش معرفی می‌شوند خلاصه‌وار در جدول 5.2 آمده‌اند.

- متغیر تصادفی را می‌توان نام یک آزمایش با خروجی تصادفی در نظر گرفت. مقدار این متغیر خروجی آزمایش است.
- توزیع احتمال برای متغیر تصادفی Y احتمال $\Pr(Y = y_i)$ را که احتمال اینکه متغیر تصادفی Y مقدار y_i را داشته باشد برای تمامی y_i ها مشخص می‌کند.
- مقدار امید، یا میانگین، متغیر تصادفی Y به صورت $E[Y] = \sum_i \Pr(Y = y_i)$ تعریف می‌شود. معمولاً از نماد μ_Y برای نمایش $E[Y]$ استفاده می‌شود.
- واریانس متغیر تصادفی Y به صورت $\text{Var}(Y) = E[(Y - \mu_Y)^2]$ تعریف می‌شود. واریانس پهنای توزیع را حول میانگین بیان

¹ error bars

می‌کند.

- انحراف از معیار متغیر تصادفی Y به صورت $\sqrt{Var(Y)}$ تعریف می‌شود. از نماد σ_Y نیز برای نمایش انحراف معیار متغیر تصادفی Y استفاده می‌کنند.
- توزیع دو جمله‌ای احتمال مشاهده‌ی r بار مشاهده‌ی شیر در پرتاب n سکه‌ی مستقل را به شرط اینکه در هر پرتاب احتمال شیر آمدن p باشد را معلوم می‌کند.
- توزیع نرمال، توزیع احتمالی زنگی شکل است که توزیع احتمال بسیاری از پدیده‌های طبیعی است.
- قضیه حد مرکزی قضیه‌ای است که می‌گوید مجموع تعداد زیادی توزیع یکسان از یک متغیر تصادفی تقریباً توزیع نرمال خواهد داشت.
- از تخمین زننده متغیر تصادفی Y برای تخمین پارامتر p از مجموعه‌ای از نمونه‌ها استفاده می‌شود.
- بایاس تخمینی متغیر تصادفی Y برای تخمین زننده‌ی پارامتر p با کمیت $(E[Y]-p)$ سنجیده می‌شود. تخمین زننده‌ای بدون بایاس است که این کمیت برایش صفر باشد.
- بازه‌ی اطمینان $N\%$ برای تخمین پارامتر p بازه‌ای است که با احتمال $N\%$ ، p را در بر خواهد گرفت.

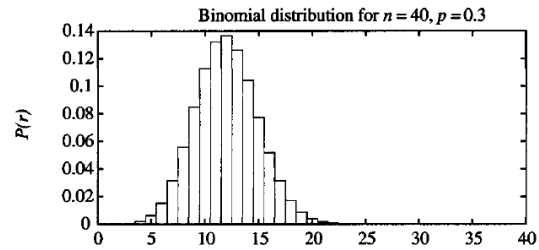
جدول 5.2 تعاریف و حقایق پایه‌ای آمار

5.3.1 تخمین خطا و تخمین ویژگی‌های توزیع دو جمله‌ای

اختلاف بین خطای واقعی و خطای نمونه‌ای دقیقاً چه رابطه‌ای با اندازه‌ی مجموعه‌ی نمونه‌ها دارد؟ این سؤال نمونه‌ای از یک مشکل آماری است: مشکل کلی خطا در تخمین ویژگی‌های کلی جامعه با داشتن مجموعه‌ای تصادفی از اعضای جامعه. در مسئله‌ی ما این ویژگی اشتباه دسته بندی شدن توسط فرضیه‌ی h است.

جواب این سؤال در توجه به این حقیقت است که اندازه گیری خطای نمونه‌ای از این طریق، آزمایشی با خروجی تصادفی است. زیرا که ابتدا مجموعه‌ی S را با n نمونه‌ی مستقل با توزیع \mathcal{D} تصادفی انتخاب می‌کنیم و خطای نمونه‌ای $error_S(h)$ را از روی این مجموعه محاسبه می‌کنیم. همان طور که در قسمت قبلی هم گفته شد، اگر آزمایش را به دفعات زیاد و هر دفعه با مجموعه‌ی S_i که n نمونه دارد تکرار کنیم، انتظار خواهیم داشت که مقادیر مختلف $error_{S_i}(h)$ مساوی نباشند. در چنین شرایطی، می‌توانیم بگوییم که $error_{S_i}(h)$ (خروجی i امین آزمایش) یک متغیر تصادفی است. در کل، می‌توان به متغیر تصادفی به چشم آزمایشی با خروجی تصادفی نگاه کرد. مقدار متغیر تصادفی نتیجه‌ی مشاهده شده‌ی آزمایش تصادفی است.

فرض کنید که می‌خواهیم k آزمایش تصادفی برای اندازه گیری متغیرهای تصادفی $error_{S_1}(h), error_{S_2}(h), \dots, error_k(h)$ انجام دهیم. و فرض کنید خروجی این آزمایش‌ها را در نموداری مستطیلی و بر اساس تعداد تکرار میزان خطا رسم می‌کنیم. با افزایش مقدار k ، نمودار مذکور به نمودار توزیع جدول 5.3 نزدیک خواهد شد. این نمودار توزیع احتمال خاصی به نام توزیع دو جمله‌ای را نشان می‌دهد.



توزیع احتمال دو جمله ای برای r بار شیر آمدن در آزمایشی که n سکه ای مجزا پرتاب می شوند، احتمال شیر آمدن در هر کدام از سکه ها p است. تابع توزیع به شکل زیر تعریف می شود:

$$P(r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$$

اگر متغیر تصادفی X از توزیع دو جمله ای پیروی کند:

- احتمال اینکه $Pr(X=r)$ مقدار r را بگیرد) با $P(r)$ مشخص می شود.

- امید X ، یا همان میانگین X ، $E[X]$ از رابطه ی زیر بدست خواهد آمد:

$$E[X] = np$$

- واریانس X ، $Var(X)$ از رابطه ی زیر بدست خواهد آمد:

$$Var(X) = np(1-p)$$

- انحراف معیار X ، σ_X از رابطه ی زیر بدست خواهد آمد:

$$\sigma_X = \sqrt{np(1-p)}$$

برای n های به اندازه ی کافی بزرگ توزیع دو جمله ای تقریباً نزدیک به توزیع نرمال با همان واریانس و میانگین خواهد بود (جدول 5.4). توصیه می شود که فقط زمانی که $np(1-p) \geq 5$ باشد از این تقریب استفاده کرد.

جدول 5.3 توزیع دو جمله ای

5.3.2 توزیع دو جمله ای

یکی از روش های خوب درک یادگیری توزیع دو جمله ای بررسی مسئله ی ذیل است. سکه ای معیوب (کج) به شما داده می شود و از شما خواسته می شود تا احتمال اینکه سکه پس از پرتاب شیر بیاید را حساب کنید. بیایید احتمال شیر آمدن سکه ی معیوب را با p نشان دهیم. شما سکه را n بار پرتاب می کنید، از این n بار شیر می آید. یک تخمین منطقی از p مقدار r/n است. توجه دارید که اگر این آزمایش را تکرار کنیم و n بار سکه را پرتاب کنیم، انتظار نمی رود که تعداد شیر های آمده دقیقاً r قبلی باشد، پس بنابراین مقدار p بدست آمده نیز با مقدار p قبلی یکی نخواهد بود. توزیع دو جمله ای احتمال وقوع مقدار r (بین 0 تا n) را در n پرتاب مشخص می کند، این احتمال با فرض اینکه تمامی پرتاب ها مستقل و احتمال شیر آمدن دقیقاً p است محاسبه می شود.

جالب است که بدانیم، تخمین p از یک دسته پرتاب مشابه تخمین $error_D(h)$ بر اساس یک دسته نمونه است. شیر بودن یک پرتاب مشابه اشتباه دسته بندی شدن یک نمونه تصادفی (با توزیع D) است. احتمال p یا همان احتمال شیر آمدن یک پرتاب مشابه احتمال اشتباه دسته بندی شدن یک نمونه تصادفی (یا همان $error_D(h)$) است. تعداد r یا همان تعداد شیرها در n پرتاب نیز مشابه تعداد دسته بندی های اشتباه نمونه ها از n نمونه تصادفی است. بنابراین r/n مشابه $error_S(h)$ است و مسئله ی تخمین p در سکه نیز مشابه مسئله ی تخمین $error_D(h)$ در فرضیه ها است. توزیع دو جمله ای فرم کلی توزیع احتمال را برای متغیر تصادفی r مشخص می کند حال فرقی نمی کند که این r تعداد شیرها باشد یا تعداد دسته بندی های اشتباه. فرم دقیق تر توزیع دو جمله ای به تعداد نمونه ها و p (یا همان $error_D(h)$) وابسته است.

شرایطی که توزیع دو جمله ای در آن صادق است:

1. مبنای کار یک آزمایش (مثل پرتاب سکه) است که خروجی اش به عنوان متغیر تصادفی (مثل Y) است. مقدار تصادفی Y می تواند فقط دو مقدار داشته باشد (برای مثال $Y=1$ برای شیر و $Y=0$ برای خط)
2. احتمال اینکه $Y=1$ شود در هر آزمایش مبنای به طور مستقل مقدار ثابت p است. پس بنابراین احتمال $Y=0$ $(1-p)$ خواهد بود. معمولاً p مجهول است و هدف یافتن تخمینی از p است.
3. سری ای از آزمایش های مبنای پشت سر هم انجام می شود (مثل پرتاب های سکه) و سری ای از متغیر های تصادفی هم ارزش مثل Y_1, Y_2, \dots, Y_n را ایجاد می کند. اگر R تعداد آزمایش ها که در آن ها $Y_i=1$ است در نظر بگیریم خواهیم داشت:

$$R \equiv \sum_{i=1}^n Y_i$$

4. احتمال اینکه متغیر تصادفی R مقدار r باشد (احتمال اینکه دقیقاً r بار شیر بیاید) بر اساس توزیع دو جمله ای به صورت زیر است:

$$\Pr(R = r) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r} \quad (5.2)$$

نموداری از این رابطه در جدول 5.3 آمده بود.

توزیع دو جمله ای احتمال تعداد خطای r در میان n نمونه را مشابه احتمال r بار شیر آمدن در میان n آزمایش پرتاب سکه توصیف می کند.

5.3.3 میانگین و واریانس

دو خاصیتی که معمولاً در مورد متغیر های تصادفی مطرح می شود امید (مقدار انتظاری یا میانگین) و واریانس است. مقدار امید، میانگین مقادیر تصادفی بدست آمده بعد از آزمایش های بسیار است. به عبارت دقیق تر:

تعریف: اگر Y یک متغیر تصادفی باشد که مقادیر y_1, \dots, y_n را بپذیرد امید $E[Y]$ ، Y به صورت زیر تعریف می شود:

$$E[Y] \equiv \sum_{i=1}^n y_i \Pr(Y = y_i) \quad (5.3)$$

برای مثال اگر متغیر تصادفی Y مقدار 1 را با احتمال 0.7 و مقدار 2 را با احتمال 0.3 بپذیرد مقدار امید $(1 \cdot 0.7 + 2 \cdot 0.3 = 1.3)$ خواهد بود. در توزیع دو جمله ای این تعریف به شکل زیر تغییر شکل پیدا می‌کند:

$$E[Y] = np \quad (5.4)$$

در این رابطه n و p پارامترهای توزیع دو جمله ای در رابطه‌ی 5.2 هستند.

کمیت دوم مطرح "پهنای" یا "میزان پخشی"² توزیع احتمال است و میزان دور بودن احتمالی متغیر تصادفی از میانگین را نشان می‌دهد.

تعریف: واریانس یک متغیر تصادفی Y ، $Var[Y]$ به فرم زیر تعریف می‌شود:

$$Var[Y] \equiv E[(Y - E[Y])^2] \quad (5.5)$$

واریانس مجموع مربعات خطای انتظاری را با استفاده از امید Y ، $E[Y]$ ، پیدا می‌کند. جزر واریانس را انحراف معیار Y می‌نامند و با σ_Y نشان می‌دهند.

تعریف: انحراف معیار متغیر تصادفی Y ، σ_Y ، به صورت زیر تعریف می‌شود:

$$\sigma_Y \equiv \sqrt{E[(Y - E[Y])^2]} \quad (5.5)$$

در شرایطی که متغیر تصادفی Y توزیع دو جمله ای داشته باشد، واریانس و انحراف از معیار به فرم زیر خواهند بود:

$$Var[Y] = np(1 - p)$$

$$\sigma_Y = \sqrt{np(1 - p)} \quad (5.7)$$

5.3.4 تخمین زنده‌ها، بایاس و واریانس

حال که نشان داده‌ایم که متغیر تصادفی $error_S(h)$ از توزیع دو جمله ای پیروی می‌کند، به سؤال اصلی بر می‌گردیم: فرق خطای نمونه ای و خطای واقعی چیست؟

بیا بید $error_S(h)$ و $error_D(h)$ را با استفاده از رابطه‌ی 5.2 که توزیع دو جمله ای را بیان می‌کند توصیف کنیم. داریم که

$$error_S(h) = \frac{r}{n}$$

$$error_D(h) = p$$

¹ width
² spread

در این رابطه n تعداد نمونه های مجموعه ی S و r تعداد دسته بندی های اشتباه h از مجموعه ی S است و p نیز احتمال دسته بندی اشتباه h از نمونه ای انتخاب شده با توزیع \mathcal{D} است.

متخصصان $error_S(h)$ را تخمین زننده ای¹ از خطای واقعی $error_D(h)$ می نامند. در کل، تخمین زننده ی یک مقدار تصادفی برای تخمین ویژگی های جمعیت آن متغیر تصادفی به کار می رود. اولین سوآلی که درباره ی هر تخمین زننده مطرح می شود این است که آیا تخمین زننده در میانگین تخمین درستی به ما می دهد؟ بایاس تخمین را به عنوان اختلاف بین مقدار امید تخمین زننده و مقدار واقعی متغیر تصادفی تعریف می کنیم.

تعریف: بایاس تخمین² برای تخمین زننده ی Y از پارامتر p به صورت

$$E[Y] - p$$

تعریف می شود.

اگر مقدار بایاس تخمین زننده صفر باشد می گوئیم که Y یک تخمین زننده ی بدون بایاس از p است. توجه داشته این حالتی است که پس از تعداد زیادی آزمایش تصادفی میانگین مقدار تصادفی به امید تخمین زننده میل کند.

آیا $error_S(h)$ تخمین زننده ی بدون بایاسی از $error_D(h)$ است؟ بله، زیرا که مقدار امید r در توزیع دو جمله ای np است (رابطه ی 5.4). حال چون که n ثابت است، پس مقدار امید r/n همان p است.

دو نکته ی قابل توجه در بایاس تخمینی وجود دارد. اول، همان طور که در ابتدای این فصل نیز گفته شد، بررسی فرضیه ها بر روی نمونه های آموزشی، تخمینی بایاس دار از خطای فرضیه به ما می دهد، این دقیقاً همان نکته ای است که بایاس تخمین به آن اشاره می کند. برای اینکه $error_S(h)$ تخمینی بدون بایاس از $error_D(h)$ به ما بدهد، باید فرضیه ی h و نمونه های S باید مستقل باشند. دوم اینکه این مفهوم نباید با بایاس استقرایی که در فصل 2 بیان شد اشتباه گرفته شود. بایاس تخمینی یک مقدار عددی است در حالی که بایاس استقرایی دسته ای از پیش فرض ها³ است.

ویژگی مهم دیگر هر تخمین زننده مقدار واریانس آن است. با داشتن انتخاب بین تخمین زننده های بدون بایاس مختلف، قابل درک است که تخمین زننده ای را انتخاب کنیم که کمترین مقدار واریانس را داشته باشد. با تعریفی که از واریانس ارائه شد، این انتخاب باعث می شود که خطای انتظاری بین تخمین و مقدار واقعی به کمترین مقدار برسد.

برای تصور این مفاهیم، فرض کنید که می خواهیم فرضیه ای را بررسی کنیم که $r=12$ خطا بر روی نمونه هایی با تعداد $n=40$ دارد. اگر $error_S(h)$ یک تخمین زننده ی بدون بایاس از $error_D(h)$ باشد، و داشته باشیم $error_S(h) = \frac{r}{n} = 0.3$. واریانس این تخمین مستقیماً به واریانس مقدار r وابسته است، زیرا که n عددی ثابت است. حال چون r با توزیع دو جمله ای انتخاب می شود برای واریانس از رابطه ی 5.7 داریم: $np(1-p)$. متأسفانه هنوز مقدار p مجهول است، اما می توان بجای آن از تخمین r/n مان از p استفاده کنیم. پس

¹ estimator

² estimation bias

³ assertion

واریانس r خواهد بود $40 \cdot 0.3(1-0.3)=8.4$ پس مقدار انحراف معیار $\sqrt{8.4} \approx 2.9$. پس انحراف معیار r/n ، $2.9/40=0.07$ خواهد بود. به طور خلاصه، $error_S(h)$ در این مثال مقدار امید 0.30 با انحراف معیار تقریباً 0.07 است (تمرین 5.1).

در کل، با داشتن خطای r در n نمونه‌ی موجود مستقل، از رابطه‌ی زیر بدست می‌آید:

$$\sigma_{error_S(h)} = \frac{\sigma_r}{n} = \sqrt{\frac{p(1-p)}{n}} \quad (5.8)$$

که با تخمین $r/n=error_S(h)$ برای p به رابطه‌ی زیر تبدیل می‌شود:

$$\sigma_{error_S(h)} = \sqrt{\frac{error_S(h)(1-error_S(h))}{n}} \quad (5.9)$$

5.3.5 بازه‌ی اطمینان

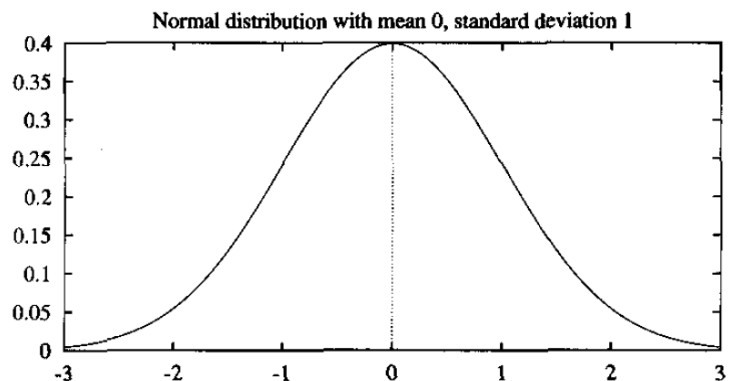
یکی از راه‌های معمول توصیف عدم قطعیت یک تخمین توجه به بازه و احتمالی است که انتظار می‌رود که مقدار واقعی در این بازه باشد است. چنین تخمینی، تخمین بازه‌ی اطمینان نامیده می‌شود.

تعریف: بازه‌ی اطمینان $N\%$ برای پارامتر p بازه‌ای است که $N\%$ احتمال می‌رود که شامل p باشد.

برای مثال، اگر مثل مثال بالا $r=12$ و $n=40$ باشد و نمونه‌ها نیز از فرضیه مستقل باشند، می‌توان گفت که با احتمال 95% مقدار $error_D(h)$ در بازه‌ی 0.30 ± 0.14 است.

بازه‌های اطمینان $error_D(h)$ چگونه بدست می‌آیند؟ جواب در این حقیقت نهفته است که توزیع احتمال دو جمله‌ای بر $error_S(h)$ حاکم است. میانگین این توزیع مقدار $error_D(h)$ است و انحراف معیار نیز از رابطه‌ی 5.9 بدست می‌آید. بنابراین، برای بدست آوردن بازه‌ی 95% فقط نیاز است که بازه را حول مقدار میانگین $error_D(h)$ بگیریم تا 95% از کل احتمال را در بر بگیرد. این بازه، بازه‌ای حول $error_D(h)$ که در 95% موارد $error_S(h)$ درون آن قرار می‌گیرد.

برای عدد معلوم N چگونه می‌توان اندازه‌ی بازه‌ی $N\%$ احتمال را بدست آورد؟ متأسفانه، این محاسبه برای توزیع احتمال دو جمله‌ای زمان‌بر است. خوشبختانه، با وجود زمان‌بری، در اکثر موارد تقریب خوبی برای بازه بدست می‌آید، زیرا که با تعداد نسبتاً زیاد نمونه توزیع به توزیع نرمال میل می‌کند. توزیع نرمال (که در جدول 5.4 نیز آمده) شاید خوش‌تعریف‌ترین توزیع احتمال باشد. همان‌طور که در جدول 5.4 نیز نشان داده شده، توزیع نرمال، توزیعی زنگی شکل حول میانگین μ و با انحراف معیار σ است. زمانی که تعداد n نسبتاً زیاد باشد، توزیع دو جمله‌ای به توزیع نرمالی با همان میانگین و انحراف معیار میل می‌کند.



توزیع نرمال (یا توزیع گوس)، توزیعی زنگی شکل است که توسط رابطه‌ی زیر تعریف می‌شود:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

اگر متغیر تصادفی X از توزیع نرمال پیروی کند:

- احتمال اینکه X در بازه‌ی (a, b) باشد از رابطه‌ی زیر بدست خواهد آمد:

$$\int_a^b p(x) dx$$

- امید یا میانگین X ، $E[X]$:

$$E[X] = \mu$$

- واریانس X ، $Var(X)$:

$$Var(X) = \sigma^2$$

- انحراف معیار X ، σ_X :

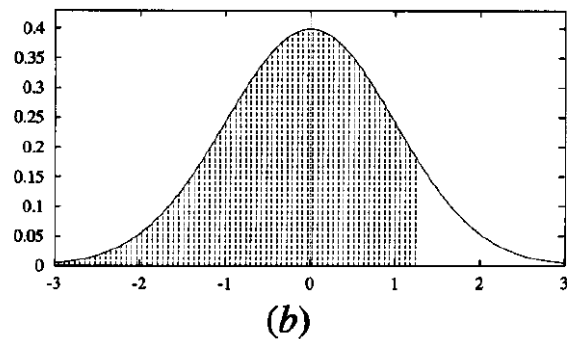
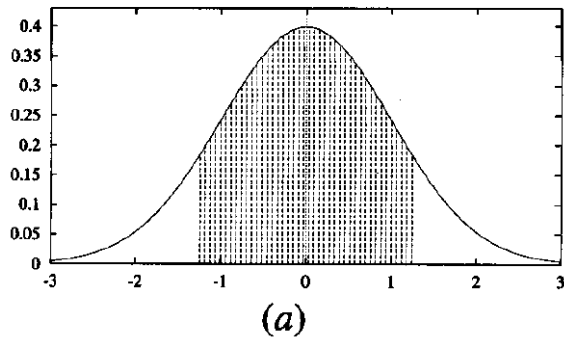
$$\sigma_X = \sigma$$

قضیه‌ی حد مرکزی¹ (در بخش 5.4.1) نشان می‌دهد که مجموع متغیرهای تصادفی را با توزیع دلخواه می‌توان با توزیع نرمال بررسی کرد.

جدول 5.4 توزیع نرمال یا توزیع گوس.

یکی از دلایلی ترجیح توزیع نرمال این است که می‌توان به راحتی بازه‌ای که $N\%$ احتمال را در بر می‌گیرد پیدا کرد. این بازه دقیقاً همان بازه‌ی $N\%$ احتمال ماست و در عمل جدول 5.1 نیز از این حقیقت بدست آمده. ثابت Z_N که در جدول 5.1 آمده بود، نصف پهنای بازه‌ی اطمینان است (فاصله‌ی بین میانگین و یکی از طرفین بازه) که بر مقدار انحراف معیار تقسیم می‌شود. شکل (a) 5.1 این بازه را برای $Z_{.80}$ نشان می‌دهد.

¹ central limit



شکل 5.1 توزیع نرمال با میانگین 0 و انحراف معیار 1.

(a) با احتمال 80٪ متغیر تصادفی در بازه‌ی از دو طرف محدود $[-1.28, 1.28]$ قرار می‌گیرد. توجه داشته باشید که $z_{.80} = 1.28$. پس با احتمال 10٪ متغیر تصادفی در سمت راست و با احتمال 10٪ متغیر تصادفی در سمت چپ این بازه قرار می‌گیرد. (b) با احتمال 90٪ متغیر تصادفی در بازه‌ی از یک طرف محدود $[-\infty, 1.28]$ قرار می‌گیرد.

به طور خلاصه، اگر متغیر تصادفی Y از توزیع نرمال با میانگین μ و انحراف معیار σ پیروی کند، مقدار تصادفی y برای Y به احتمال $N\%$ در بازه‌ی زیر قرار می‌گیرد:

$$\mu \pm z_N \sigma \quad (5.10)$$

به طور مشابه، مقدار میانگین μ با احتمال $N\%$ در بازه‌ی زیر قرار می‌گیرد:

$$y \pm z_N \sigma \quad (5.11)$$

این واقعیت را می‌توان به سادگی با واقعیت‌های کلی قبلی ذکر شده در مورد بازه‌ی $N\%$ در توابع گسسته مقدار ترکیب کرد (رابطه‌ی 5.1). ابتدا اینکه می‌دانیم $error_S(h)$ از توزیع دو جمله‌ای پیروی می‌کند که میانگین آن $error_D(h)$ است و انحراف معیارش نیز از رابطه‌ی 5.9 بدست می‌آید. دوم اینکه می‌دانیم که برای زمانی که تعداد n به اندازه‌ی کافی بزرگ باشد توزیع دو جمله‌ای را می‌توان با تقریب خوبی با توزیع نرمال تقریب زد. سوم اینکه رابطه‌ی 5.11 روش پیدا کردن بازه‌ی اطمینان $N\%$ را با توجه به توزیع نرمال مشخص می‌کند. بنابراین، با جایگزینی میانگین و انحراف معیار $error_S(h)$ در رابطه‌ی 5.11 برای توابع گسسته مقدار به رابطه‌ی 5.1 می‌رسیم.

$$error_S(h) \pm z_N \sqrt{\frac{error_S(h)(1 - error_S(h))}{n}}$$

دو نکته‌ی مهم در تخمین این رابطه به شرح زیرند:

1. در تخمین انحراف معیار σ برای $error_S(h)$ ، ما از $error_S(h)$ به جای $error_D(h)$ استفاده کردیم (در نتیجه گیری رابطه‌ی 5.9 از رابطه‌ی 5.8)

2. توزیع دو جمله‌ای را با توزیع نرمال تخمین زده‌ایم.

این دو تقریب تا زمانی که $n \geq 30$ و $np(1-p) \geq 5$ تقریب‌های خوبی هستند. برای مقادیر کمتر n بهتر است از جدولی با مقادیر توزیع دو جمله‌ای به جای توزیع نرمال استفاده کنیم.

5.3.6 بازه های یک طرفه و دو طرفه

توجه دارید که بازه‌ی اطمینان ذکر شده مرز دوطرفه¹ دارد؛ زیرا که کمیت تخمین زده شده را هم از بالا و هم از پایین محدود کرده است. در بعضی موارد، علاقه‌ی ما فقط به یکی از این دو مرز است (مرز یک طرفه²). برای مثال ممکن است جواب سؤال "احتمال اینکه $error_D(h)$ حداقل از U بیشتر باشد؟" برایمان مهم باشد. چنین سؤال‌هایی که مرز یک طرفه دارند طبیعتاً زمانی ایجاد می‌شوند که ما به محدود کردن حداکثر خطای h علاقه داریم و اینکه خطا از آن مقدار بسیار کوچک‌تر باشد برایمان مهم نیست.

تغییر کوچکی در فرایند بالا آن‌را به روش پیدا کردن مرز یک طرفه تبدیل می‌کند. نکته‌ی اساسی این تغییر این است که توزیع نرمال حول میانگینش متقارن پخش شده است. به همین خاطر می‌توان هر بازه‌ی اطمینان با مرز دوطرفه را به بازه‌ی h با مرز یک طرفه تبدیل کرد (مثل شکل (b) 5.1). اگر بازه‌ی اولیه اطمینان $100(1-\alpha)\%$ داشته باشد، بازه‌ی یک طرفه‌ی فقط محدود از بالا اطمینان $100(1-\alpha/2)\%$ خواهند داشت. بازه‌ی یک طرفه‌ی فقط محدود از پایین نیز همین اطمینان را خواهد داشت. در اینجا α احتمال این است که متغیر تصادفی خارج بازه‌ی مزبور باشد. به عبارت دیگر، α احتمال این است که متغیر تصادفی در قسمت‌هایی از شکل (a) 5.1 قرار بگیرد که هاشور نخورده‌اند. متناسباً مقدار $\alpha/2$ نیز احتمال این است که متغیر تصادفی در قسمت هاشور نخورده‌ی شکل (b) 5.1 قرار بگیرد.

برای تصور، دوباره فرض کنید که فرضیه‌ی $r=12$ خطا بر روی مجموعه‌ی $n=40$ نمونه داریم که نمونه‌ها از فرضیه مستقلند. همان طور که بالاتر نیز گفته شد، این اطلاعات بازه‌ی 95% (دوطرفه) 0.30 ± 0.14 را مشخص می‌کند. در اینجا، $100(1-\alpha)=95\%$ پس $\alpha=0.05$. بنابراین بدون اضافه کردن هیچ پیش فرض اضافه‌ای، می‌توانیم بگوییم که با اطمینان $100(1-\alpha/2)=97.5\%$ ، $error_D(h)$ حداکثر $0.30+0.14=.44$ است. بنابراین، مرزی یک طرفه برای $error_D(h)$ با اطمینان دو برابر نسبت به مرز دوطرفه خواهیم داشت (تمرین 5.3).

5.4 روش کلی برای استخراج بازه های اطمینان

در قسمت قبل چگونگی بدست آوردن بازه های تخمین برای حالت خاص: تخمین $error_D(h)$ برای توابع گسسته مقدار h بر پایه‌ی مجموعه‌ی n نمونه توضیح داده شد. روشی که آنجا ارائه شد روشی کلی‌تر برای استفاده در تعداد کثیری از مشکلات تخمین شرح می‌دهد. در کل، می‌توان مسئله‌ی فوق را تخمین میانگین (مقدار امید) یک مجموعه بر اساس زیر مجموعه‌ی n عضو دانست. فرایند کلی مراحل زیر را در بر می‌گیرد:

1. معلوم کردن پارامتر p ای که از مجموعه می‌خواهیم تخمین بزنیم، برای مثال $error_D(h)$.
2. تعریف تخمین زنده‌ی Y ، مثل $error_S(h)$. بهتر است این تخمین زنده واریانس کم داشته و بدون بایاس باشد.
3. مشخص کردن توزیع احتمال D_Y که کار تخمین زنده‌ی Y را کنترل می‌کند. مشخص کردن این توزیع شامل مشخص کردن واریانس و میانگین نیز می‌شود.

¹ two sided bound

² one side bound

4. مشخص کردن بازه‌ی N% با پیدا کردن مقادیر آستانه‌ی L و U که N% از جرم احتمال توزیع D_Y بین دو مقدار L و U قرار بگیرد.

در بخش‌های بعدی این فصل، از این روش کلی برای مسائل تخمین مختلفی که در یادگیری ماشین مطرح است استفاده می‌کنیم. با این وجود، ابتدا بیاید نتیجه‌ی اساسی قضیه‌ی حد مرکزی را بررسی کنیم.

5.4.1 قضیه‌ی حد مرکزی

یکی از حقیقت‌هایی که برای پیدا کردن بازه‌ی اطمینان مورد استفاده قرار می‌گیرد قضیه‌ی حد مرکزی است. دوباره شرایط کلی، n متغیر تصادفی مستقل Y_1, \dots, Y_n که از توزیع احتمال مجهول خاصی پیروی می‌کنند (مثل n بار پرتاب یک سکه) را در نظر بگیرید. فرض کنید μ میانگین این توزیع مجهول باشد و σ نیز انحراف معیار آن باشد. این متغیرهای Y_i مستقل و به طور یکسان توزیع¹ شده‌اند زیرا که هر کدام از آن‌ها آزمایش مجزایی را توصیف می‌کند، و هر کدام از همان توزیع احتمال پیروی می‌کنند. در تخمین میانگین μ تابع توزیع حاکم بر Y_i ها از همان تعریف میانگین استفاده می‌کنیم $\bar{Y}_n \equiv \frac{1}{n} \sum_{i=1}^n Y_i$. (مثل نسبت شیرها به کل پرتاب‌ها). قضیه‌ی حد مرکزی می‌گوید که توزیع احتمال حاکم بر \bar{Y}_n مستقل از اینکه توزیع احتمال Y_i ها چه باشد با $n \rightarrow \infty$ به توزیع نرمال میل می‌کند. علاوه بر این توزیعی که \bar{Y}_n را کنترل می‌کند میانگین μ و انحراف معیار $\frac{\sigma}{\sqrt{n}}$ خواهد داشت. به عبارت دیگر،

قضیه‌ی 5.1 قضیه‌ی حد مرکزی. فرض کنید که دسته‌ی متغیر تصادفی مستقل و به طور یکسان توزیع شده‌ی Y_1, \dots, Y_n را داریم که هر کدام از این متغیرهای تصادفی با توزیع احتمالی که میانگین μ و واریانس σ^2 دارد کنترل می‌شوند. اگر $\bar{Y}_n \equiv \frac{1}{n} \sum_{i=1}^n Y_i$ تعریف کنیم، و $n \rightarrow \infty$ توزیع احتمال حاکم بر

$$\frac{\bar{Y}_n - \mu}{\frac{\sigma}{\sqrt{n}}}$$

به توزیع نرمال با میانگین صفر و انحراف معیار 1 میل خواهد کرد.

این حقیقت بسیار جالبی است، توزیع احتمال حاکم بر \bar{Y}_n بدون دانستن توزیع احتمال حاکم بر Y_i ها معلوم می‌گردد! علاوه بر آن، قضیه‌ی حد مرکزی روشی برای پیدا کردن واریانس و میانگین تابع توزیع Y_i ها از واریانس و میانگین تابع توزیع \bar{Y}_n ارائه می‌کند.

قضیه‌ی حد مرکزی، حقیقتی پر کاربرد است، زیرا که به هر صورت که یک تخمین زننده برای تخمین میانگین چیزی تعریف کنیم ($errors(h)$ تخمین زننده‌ی میانگین خطاست)، برای n های به اندازه‌ی کافی بزرگ آن‌را می‌توان با توزیع نرمال تخمین زد. حال اگر واریانس این توزیع نرمال (تخمینی) را بدانیم می‌توانیم با استفاده از رابطه‌ی 5.11 برای محاسبه‌ی بازه‌ی اطمینان استفاده کنیم. یک تقریب متداول این است که زمانی می‌توانیم از تقریب نرمال استفاده کنیم که داشت باشیم $n \geq 30$. توجه دارید که در قسمت قبلی از چنین توزیع نرمالی برای تخمین توزیع دوجمله‌ای که $errors(h)$ را توصیف می‌کرد استفاده کردیم.

¹ identically distributed

5.5 تفاوت خطاهای دو فرضیه

حالتی را تصور کنید که دو فرضیه‌ی h_1 و h_2 را برای تابع هدف گسسته مقداری داریم. فرضیه‌ی h_1 بر روی مجموعه‌ی S_1 که شامل n_1 نمونه است، و فرضیه‌ی h_2 بر روی مجموعه‌ی S_2 که شامل n_2 نمونه است بررسی شده است. فرض کنید که می‌خواهیم تفاوت بین خطای واقعی این دو فرضیه را تخمین بزنیم.

$$d \equiv \text{error}_D(h_1) - \text{error}_D(h_2)$$

در اینجا ما از فرایند چهار مرحله‌ای ارائه شده در ابتدای بخش 5.4 برای بدست آوردن بازه‌ی اطمینان برای d استفاده می‌کنیم. با معلوم کردن d به عنوان پارامتری که می‌خواهیم تخمین بزنیم، باید یک تخمین زننده معرفی کنیم. تنها انتخاب ممکن و واضح برای تخمین زننده‌ی d اختلاف بین خطاهای نمونه‌ای است که با \hat{d} نشان می‌دهیم:

$$\hat{d} \equiv \text{error}_D(h_1) - \text{error}_D(h_2)$$

با وجود اینکه اینجا اثبات نمی‌کنیم اما می‌توان نشان داد که \hat{d} تخمینی بدون بایاس از d ارائه می‌کند: $E[\hat{d}] = d$.

اما \hat{d} از چه توزیع احتمالی پیروی می‌کند؟ با استفاده از آنچه در قسمت‌های قبلی گفته شد، اگر n_1 و n_2 به اندازه‌ی کافی بزرگ باشند (هر دو بزرگ‌تر از 30 باشند) هر دو متغیر تصادفی $\text{error}_{S_1}(h_1)$ و $\text{error}_{S_2}(h_2)$ از توزیع نرمال پیروی خواهند کرد. چون تفاوت دو توزیع نرمال نیز توزیعی نرمال است، \hat{d} نیز توزیعی نرمال با میانگین d خواهد داشت. همچنین می‌توان نشان داد که واریانس این توزیع مجموع واریانس توزیع‌های $\text{error}_{S_1}(h_1)$ و $\text{error}_{S_2}(h_2)$ است. با استفاده از رابطه‌ی 5.9 برای تخمین واریانس هر یک از این دو توزیع داریم که:

$$\sigma_{\hat{d}}^2 \approx \frac{\text{error}_{S_1}(h_1) (1 - \text{error}_{S_1}(h_1))}{n_1} + \frac{\text{error}_{S_2}(h_2) (1 - \text{error}_{S_2}(h_2))}{n_2} \quad (5.12)$$

حال که توزیع احتمالی که \hat{d} را کنترل می‌کند را مشخص کرده‌ایم، به راحتی می‌توان بازه‌ی اطمینان را که از \hat{d} برای d بدست می‌آید مشخص کرد. برای متغیر تصادفی \hat{d} که از توزیع نرمالی با میانگین d و واریانس σ^2 پیروی می‌کند بازه‌ی اطمینان $\hat{d} \pm z_N \sigma$ را خواهیم داشت. با استفاده از واریانس تخمینی $\sigma_{\hat{d}}^2$ که در بالا محاسبه شد این بازه‌ی اطمینان تخمینی برای d به صورت زیر خواهد بود:

$$\hat{d} \pm z_N \sqrt{\frac{\text{error}_{S_1}(h_1) (1 - \text{error}_{S_1}(h_1))}{n_1} + \frac{\text{error}_{S_2}(h_2) (1 - \text{error}_{S_2}(h_2))}{n_2}} \quad (5.13)$$

در این رابطه z_N مقادیری است که از جدول 5.1 استخراج می‌شود. رابطه‌ی بالا بازه‌ی اطمینان دوطرفه‌ای برای تخمین اختلاف بین خطاهای دو فرضیه به ما می‌دهد. بعضی مواقع ممکن است علاقه‌ی ما به بازه‌ی یک طرفه باشد، محدود کردن بزرگ‌ترین اختلاف خطاها در یک محدوده‌ی خاص. این بازه‌ی اطمینان دوطرفه را می‌توان با همان فرایند قسمت 5.3.6 به بازه‌ی یک طرفه تبدیل کرد.

با این وجود که بررسی بالا در حالتی انجام گرفته که دو فرضیه‌ی h_1 و h_2 روی مجموعه داده‌های مستقل تست شده‌اند، اما گاهی استفاده از این بازه‌ی اطمینان (رابطه‌ی 5.13) در جایی که h_1 و h_2 هر دو بر روی مجموعه‌ی S (که هنوز از هر دو فرضیه‌ی h_1 و h_2 مستقل است) قابل قبول است. در این حالت می‌توان \hat{d} را به فرم زیر تعریف کرد:

$$\hat{d} \equiv error_S(h_1) - error_S(h_2)$$

این اختلاف در \hat{d} جدید معمولاً کمتر از اختلاف رابطیه‌ی 5.12 است، زیرا که دو مجموعه‌ی S_1 و S_2 هر دو S در نظر گرفته شده‌اند. دلیل این کاهش استفاده از یک مجموعه‌ی نمونه‌ی S برای ارزیابی اختلاف اثر اختلافات تصادفی بین ترکیب S_1 و S_2 را حذف خواهد کرد. در این حالت، بازه‌ی اطمینان رابطیه‌ی 5.13 در کل بازه‌ی محافظه کارانه، اما هنوز درست، خواهد بود.

5.5.1 تست فرضیه¹

بعضی مواقع، علاقه‌ی ما بیشتر به احتمال درستی یک حدس است تا اینکه بازه‌های اطمینان را برای پارامترهای حدس داشته باشیم. برای مثال، فرض کنید، علاقه‌ی ما به این سؤال که "با چه احتمالی داریم $error_D(h_1) > error_D(h_2)$ ؟" است. با توجه به آنچه در قسمت‌های گذشته گفتیم، فرض کنید که دو فرضیه‌ی h_1 و h_2 را بر روی دو مجموعه‌ی مستقل S_1 و S_2 با اندازه‌ی مساوی 100 تست می‌کنیم و داریم، $error_{S_1}(h_1) = .30$ و $error_{S_2}(h_2) = .20$ ، بنابراین اختلاف مشاهده شده $\hat{d} = .10$ خواهد بود. البته با توجه به اختلافات تصادفی در داده‌های نمونه‌ای ممکن است چنین نتایجی حتی زمانی که $error_D(h_1) \leq error_D(h_2)$ است نیز مشاهده شود. احتمال اینکه داشته باشیم $error_D(h_1) > error_D(h_2)$ با داشتن اینکه اختلاف نمونه‌ای $\hat{d} = .10$ را داریم چقدر است؟ یا به طور معادل احتمال اینکه $d > 0$ باشد به شرط اینکه $\hat{d} = .10$ چقدر است؟

توجه دارید که احتمال اینکه $Pr(d > 0)$ مشابه احتمال این است که \hat{d}, d را به اندازه‌ی 10. بیشتر تخمین زده باشد. به عبارت دیگر، احتمال این است که \hat{d} در بازه‌ی اطمینان یک طرفه‌ی 10. $\hat{d} < d +$ قرار بگیرد است. در این رابطه d میانگین توزیع احتمال حاکم بر \hat{d} است پس می‌توان رابطه را به صورت $\hat{d} < \mu_{\hat{d}} + 10.$ بازنویسی کرد.

به طور خلاصه احتمال $Pr(d > 0)$ مساوی این احتمال است که \hat{d} در بازه‌ی اطمینان یک طرفه‌ی 10. $\hat{d} < \mu_{\hat{d}} +$ قرار داشته باشد است. از آنجایی که در قسمت قبلی توزیع احتمال حاکم بر \hat{d} را محاسبه کرده‌ایم، می‌توان احتمال اینکه \hat{d} در بازه‌ی اطمینان یک طرفه‌ی قرار گیرد با جرم احتمال توزیع \hat{d} در این بازه اندازه گیری خواهد شد.

بیابید این محاسبه را با بازنویسی دوباره‌ی $\hat{d} < \mu_{\hat{d}} + 10.$ با انحراف از معیار شروع کنیم. با استفاده از رابطیه‌ی 5.12 می‌توان بدست آورد که $\sigma_{\hat{d}} \approx .061$ ، پس می‌توان بازه‌ی اطمینان را به فرم زیر نوشت،

$$\hat{d} < \mu_{\hat{d}} + 1.64\sigma_{\hat{d}}$$

میزان احتمال متناسب با این بازه‌ی یک طرفه در توزیع نرمال چند است؟ با توجه به جدول 5.1، می‌توان بدست آورد که بازه‌ی 1.64 برابر انحراف حول میانگین برای بازه‌ی دوطرفه احتمال 90٪ دارد. بنابراین، بازه‌ی اطمینان دو طرفه احتمال 95٪ را خواهد داشت.

¹ Hypothesis testing

بنابراین، با داشتن مشاهده‌ی $\hat{d} = 0.10$ ، احتمال اینکه $error_D(h_1) > error_D(h_2)$ تقریباً 95٪ است. در واژگان ادبیات آماری، می‌گوییم که این فرضیه که " $error_D(h_1) > error_D(h_2)$ " را با اطمینان 95٪ می‌پذیریم. یا به طور مشابه ممکن است بگوییم که فرضیه متضاد¹ (یا فرضیه‌ی تهی²) را با احتمال $0.05 = (1 - 0.95)$ رد می‌کنیم.

5.6 مقایسه‌ی الگوریتم‌های یادگیری

بعضی مواقع مقایسه‌ی عملکرد دو الگوریتم یادگیری L_A و L_B برای ما از مقایسه‌ی دو فرضیه اهمیت بیشتری دارد. آزمون مناسب برای مقایسه‌ی الگوریتم‌های یادگیری چیست و چگونه می‌توان معلوم کرد که تفاوت‌های بدست آمده از نظر آماری قابل توجهند؟ با وجود اینکه بحث‌ها هنوز در این مبحث از یادگیری ماشین داغ است اما ما در اینجا روشی خاص را معرفی خواهیم کرد. بحث در مورد دیگر متد‌های جایگزین را می‌توانید در (Ditterich 1996) پیدا کنید.

مثل قبل، کار را با تعیین پارامتری که می‌خواهیم تخمین بزنیم آغاز می‌کنیم. فرض کنید قصد داریم مشخص کنیم که کدام یک از دو الگوریتم L_A یا L_B به طور متوسط برای یادگیری تابع هدف f معلوم متناسب‌ترند. یکی از راه‌های تعریف "به طور متوسط" بررسی کارایی نسبی دو الگوریتم بر روی تمامی مجموعه نمونه‌های n عضوی ممکن که با استفاده از توزیع احتمال نمونه‌ای D است. به عبارت دیگر، قصد داریم که مقدار امید خطای بین دو فرضیه را تخمین بزنیم

$$E_{S \subset D} [error_D(L_A(S)) - error_D(L_B(S))] \quad (5.14)$$

در این رابطه $L(S)$ فرضیه‌ای است که با استفاده از متد L از نمونه‌های S بدست می‌آید، $S \subset D$ نیز به این معناست که مقدار امید بر روی نمونه‌های S که با توزیع D انتخاب می‌شوند محاسبه می‌شود. عبارت بالا مقدار امید اختلاف بین خطاهای یادگیری دو متد L_A و L_B را نشان می‌دهد.

البته در عمل مجموعه‌ای محدود از نمونه‌ها D_0 در دسترس است و بررسی بین دو متد را بر روی این مجموعه‌ی محدود انجام می‌دهیم. در چنین شرایطی، یکی از روش‌های ساده‌ی تخمین کمیت بالا تقسیم D_0 به دسته‌ی آموزشی S_0 و دسته‌ی تست T_0 است. داده‌های آموزشی را می‌توان برای آموزش در هر دو روش L_A و L_B به کار برد و از دسته‌ی تست می‌توان برای مقایسه‌ی دقت هر کدام از فرضیه‌های یاد گرفته شده استفاده کرد. به عبارت دیگر، ما کمیت زیر را محاسبه می‌کنیم:

$$error_{T_0}(L_A(S_0)) - error_{T_0}(L_B(S_0)) \quad (5.15)$$

توجه داشته باشید که این تخمین زننده و کمیت رابطه‌ی 5.14 دو تفاوت کلیدی دارند. ابتدا اینکه در این رابطه از $error_{T_0}(h)$ برای تخمین مقدار $error_D(h)$ استفاده شده است. دوم اینکه در این رابطه فقط تفاوت بین خطاها برای یک مجموعه‌ی آموزشی S_0 به جای کل مجموعه‌های آموزشی ممکن توزیع D استفاده شده است.

¹ opposite hypothesis

² null hypothesis

یکی از راه های بهبود این تخمین زنده‌ی رابطه‌ی 5.15 تقسیم داده های در چندین مرحله D_0 به مجموعه های کوچک تر و استفاده از میانگین خطای بدست آمده از دسته‌ی تست در آزمایش های مختلف است. این کار به فرایند نشان داده شده در جدول 5.5 برای مقایسه‌ی خطاهای دو متد یادگیری بر اساس داده های ثابت D_0 می انجامد. این فرایند تقسیم ابتدا داده ها را به k زیر مجموعه‌ی هم اندازه که هر کدام حداقل 30 نمونه دارند تقسیم می کند. سپس الگوریتم یادگیری را k بار آموزش می دهد و آزمایش می کند، در هر یک از این k بار یکی از k زیر مجموعه به عنوان مجموعه‌ی تست مورد استفاده قرار می گیرد و بقیه داده ها نیز مجموعه‌ی آموزشی خواهند بود. به این ترتیب، الگوریتم های یادگیری بر روی k مجموعه‌ی مجزای تست بررسی می شوند و میانگین اختلاف در خطا های $\bar{\delta}$ به عنوان یک تخمین زنده برای اختلاف بین دو الگوریتم یادگیری انتخاب می شود.

کمیت $\bar{\delta}$ که از فرایند جدول 5.5 بدست می آید را می توان به عنوان تخمینی از کمیت مطلوب رابطه‌ی 5.14 به حساب آورد. حتی می توان به $\bar{\delta}$ به دید تخمینی از کمیت زیر نگاه کرد:

$$E_{S \subset D_0} [error_D(L_A(S)) - error_D(L_B(S))] \quad (5.16)$$

در این رابطه S مجموعه ای دلخواه از $|D_0| \frac{k-1}{k}$ نمونه است که با توزیع یکنواخت از D_0 انتخاب شده اند. تنها تفاوت بین این کمیت و کمیت اصلی ما در رابطه‌ی 5.14 این است که این کمیت مقدار امید را بر روی زیرمجموعه ای از داده های موجود D_0 پیدا می کند به جای اینکه از تمامی نمونه ها با توزیع D استفاده کند.

1. داده های موجود D_0 را به k دسته‌ی مجزای T_1, T_2, \dots, T_k با اندازه های مساوی تقسیم کن، اندازه‌ی هر مجموعه باید حداقل 30 باشد.

2. برای تمامی مقادیر i با شروع از 1 و کمتر مساوی از k :
از T_i برای دسته‌ی تست و از بقیه‌ی داده ها برای دسته‌ی آموزشی S_i استفاده کن.

$$\begin{aligned} S_i &\leftarrow \{D_0 - T_i\} \\ h_A &\leftarrow L_A(S_i) \\ h_B &\leftarrow L_B(S_i) \\ \delta_i &\leftarrow error_{T_i}(h_A) - error_{T_i}(h_B) \end{aligned}$$

3. مقدار $\bar{\delta}$ را از تعریف زیر خروجی بده:

$$\bar{\delta} \equiv \frac{1}{k} \sum_{i=1}^k \delta_i \quad (T5.1)$$

جدول 5.5 فرایند تخمین تفاوت بین خطاهای بین دو متد یادگیری L_A و L_B بازه های اطمینان این تخمین در متن آورده شده اند.
بازه‌ی اطمینان $N\%$ برای تخمین کمیت رابطه‌ی 5.16 با $\bar{\delta}$ به فرم زیر است:

$$\bar{\delta} \pm t_{N,k-1} S_{\bar{\delta}} \quad (5.17)$$

در این رابطه $t_{N,k-1}$ ثابتی است که نقش Z_N را در تعریف قبلی بازه‌ی اطمینان بازی می کند، $S_{\bar{\delta}}$ نیز تخمین انحراف معیار توزیع حاکم بر $\bar{\delta}$ است. در کل، $S_{\bar{\delta}}$ به صورت زیر تعریف می شود:

$$s_{\bar{\delta}} \equiv \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^k (\delta_i - \bar{\delta})^2} \quad (5.18)$$

توجه دارید که ثابت $t_{N,k-1}$ دو اندیس دارد. اندیس اول درصد اطمینان بازه را مشخص می‌کند، مشابه Z_N . پارامتر دوم که درجه‌ی آزادی¹ نیز نامیده می‌شود و معمولاً با ν نمایش داده می‌شود، این پارامتر تعداد فرایندهای تصادفی مستقل که برای تولید مقدار تصادفی $\bar{\delta}$ انجام می‌شود را نشان می‌دهد. در شرایط حاضر، درجه‌ی آزادی همان $k-1$ است. مقادیر ثابت t در جدول 5.6 آورده شده، توجه دارید که با $k \rightarrow \infty$ مقدار $t_{N,k-1}$ به ثابت Z_N میل می‌کند.

توجه دارید که فرایندی که برای مقایسه‌ی دو متد یادگیری در اینجا آورده شد دقت هر دو فرضیه‌ی یاد گرفته شده را بر اساس یک دسته‌ی تست مشترک بررسی می‌کند. این با چیزی که در قسمت 5.5 در مورد مقایسه‌ی فرضیه‌ها با دسته تست‌های مستقل گفته شد در تضاد است. به تست‌هایی که فرضیه‌ها بر روی مجموعه‌های مشابهی تست می‌شوند تست‌های جفت² می‌گویند. تست‌های جفت معمولاً بازه‌های اطمینان کوچک‌تری ایجاد می‌کنند زیرا که در خطاهای مشاهده شده در تست‌های جفت فقط بخاطر اختلاف در فرضیه‌هاست. در مقابل، زمانی که فرضیه‌ها بر روی داده‌های مجزایی تست می‌شوند، تفاوت ناشی از ترکیب دو مجموعه‌ی نمونه ممکن است بر روی تست تأثیر بگذارد.

درجه‌ی اطمینان N				
	99%	98%	95%	90%
$\nu=2$	9.92	6.96	4.30	2.92
$\nu=5$	4.03	3.36	2.57	2.02
$\nu=10$	3.17	2.76	2.23	1.81
$\nu=20$	2.84	2.53	2.09	1.72
$\nu=30$	2.75	2.46	2.04	1.70
$\nu=120$	2.62	2.36	1.98	1.66
$\nu=\infty$	2.58	2.33	1.96	1.64

جدول 5.6 مقادیر $t_{N,\nu}$ برای بازه‌های اطمینان دو طرفه. با $\nu \rightarrow \infty$ $t_{N,\nu}$ به Z_N میل می‌کند.

¹ number of degrees of freedom

² paired tests

5.6.1 تست‌های جفتی t^1

در بالا، فرایندی برای مقایسه‌ی دو متد یادگیری با مجموعه‌ی ثابتی از داده‌ها ارائه شد. در این بخش توجیه آماری برای این فرایند و بازه‌های اطمینان روابط 5.17 و 5.18 را مورد بحث قرار می‌دهیم. در اولین بار خواندن این کتاب می‌توانید بدون از دست دادن پیوستگی مطالب این قسمت را نخوانید.

بهترین راه برای درک توجیه بازه‌های تخمینی اطمینان که در رابطه‌ی 5.17 آورده شده در نظر گرفتن مسئله‌ی تخمینی زیر است:

- دسته‌ای از مقادیر تصادفی مشاهده شده‌ی مستقل و به طور یکسان توزیع شده داریم.
- سعی داریم که میانگین μ را برای توزیع حاکم بر Y_i ها را پیدا کنیم.
- تخمین زنده‌ی مورد استفاده در این مسئله مقدار \bar{Y} است:

$$\bar{Y} = \frac{1}{k} \sum_{i=1}^k Y_i$$

این مسئله‌ی تخمین میانگین μ که بر اساس مقدار \bar{Y} صورت می‌گیرد، بسیار کلی است. برای مثال، این حالت کلی مسئله‌ی تخمین $error_D(h)$ با استفاده از $error_S(h)$ را نیز شامل می‌شود. (در این مسئله، Y_i ها 1 یا 0 بودند (بنا به اینکه نمونه درست دسته بندی می‌شود یا خیر) و $error_D(h)$ همان میانگین بود که می‌خواستیم تخمین بزنیم). تست t ، که در رابطه‌ی 5.17 و 5.18 نیز آورده شده است، حالت خاصی از این مسئله است، حالتی که Y_i ها از توزیع نرمال پیروی می‌کنند.

حال فرم ایده آل زیر را برای متد جدول 5.5 را برای مقایسه‌ی کارایی دو الگوریتم یادگیری در نظر بگیرید. فرض کنید که به جای داشتن مجموعه‌ی ثابت D_0 ، می‌توانیم نمونه‌های آموزشی جدید را بر اساس توزیع احتمال \mathcal{D} دریافت کنیم. در کل، متد ایده آل فرایند جدول 5.5 در هر بار تکرار حلقه از یک دسته‌ی آموزشی جدید S_i و دسته‌ی تست جدید T_i را که با توزیع \mathcal{D} (همان توزیع D_0) ایجاد شده استفاده می‌کند. این متد ایده آل کاملاً با فرم مسئله‌ی تخمینی بالا تطبیق دارد. در کل، معیار δ_i در این فرایند متناسب با متغیرهای تصادفی به طور یکسان توزیع شده Y_i ها هستند. میانگین μ این توزیع‌ها مقدار امید اختلاف بین خطاها بین دو متد یادگیری را نشان می‌دهد (رابطه‌ی 5.14). میانگین نمونه‌ای \bar{Y} کمیتی است δ که توسط حالت ایده آل این متد اندازه گیری می‌شود. علاقه‌ی ما به جواب این سؤال است که "میانگین δ تا چه میزان تخمین خوبی از μ به ما می‌دهد؟"

ابتدا توجه داشته باشید که اندازه‌ی مجموعه‌های تست T_i طوری انتخاب می‌شود که هر یک از مجموعه‌ها حداقل 30 نمونه داشته باشد. به همین دلیل، هر کدام از δ_i ها (طبق قضیه‌ی حد مرکزی) توزیعی تقریباً نرمال خواهند داشت. بنابراین، با حالت خاصی روبرو هستیم که در آن توزیع حاکم بر Y_i ها همگی تقریباً نرمال هستند. می‌توان نشان داد که در کل، زمانی که توزیع حاکم بر هر یک Y_i توزیعی نرمال است، توزیع حاکم بر میانگین نمونه‌ای \bar{Y} توزیعی نرمال خواهد بود. با این دانش که \bar{Y} توزیعی نرمال دارد، می‌توان از آنچه پیش‌تر در مورد بازه‌های اطمینان گفته شد برای این متغیر تصادفی استفاده کرد (رابطه‌ی 5.11 که برای توزیع احتمال‌های با توزیع نرمال صادق بود). متأسفانه، این رابطه نیاز به انحراف معیار دارد، که در حال حاضر برای ما متغیری مجهول است.

¹ Paired t tests

تست t دقیقاً برای چنین شرایطی به وجود آمده است، شرایطی که در آن هدف تخمین میانگین نمونه ای مجموعه ای از متغیرهای تصادفی با توزیع‌های نرمال به طور یکسان توزیع شده است. در چنین شرایطی، می‌توان از بازه‌های اطمینان رابطه‌ی 5.17 و 5.18 که با نماد گذاری جدید به شکل زیر نمایش داده می‌شوند استفاده کرد:

$$\mu = \bar{Y} \pm t_{N,k-1} S_{\bar{Y}}$$

که در این رابطه $S_{\bar{Y}}$ انحراف معیار میانگین نمونه است:

$$S_{\bar{Y}} = \sqrt{\frac{1}{k(k-1)} \sum_{i=1}^k (Y_i - \bar{Y})^2}$$

و $t_{N,k-1}$ ثابتی مشابه ثابت قبلی Z_N است. در واقع ثابت $t_{N,k-1}$ ثابتی است که ویژگی‌های ناحیه ای از توزیع احتمال موسوم به توزیع t را مشابه ثابت Z_N که ناحیه ای از توزیع احتمال نرمال را مشخص می‌کند. توزیع t مشابه توزیع نرمال توزیعی زنگی شکل است با این تفاوت که پهنای بیشتری دارد تا واریانس $S_{\bar{Y}}$ را داشته باشد تا بتواند انحراف معیار واقعی $\sigma_{\bar{Y}}$ را تخمین بزند. توزیع t با میل کردن متغیر k به سمت بینهایت به توزیع نرمال (و متناسباً $t_{N,k-1}$ نیز به Z_N) میل خواهد کرد. این منطقی است زیرا که انتظار داریم که $S_{\bar{Y}}$ با افزایش k به سمت مقدار واقعی انحراف معیار $\sigma_{\bar{Y}}$ میل کند و همچنین زیرا که زمانی که انحراف معیار را دقیقاً داریم می‌توانیم از Z_N استفاده کنیم.

5.6.2 نکات کاربردی

توجه دارید که بحث بالا استفاده از تخمین بازه‌ی اطمینان رابطه‌ی 5.17 را در حالتی که علاقه ما به میانگین نمونه ای \bar{Y} برای تخمین میانگین مجموعه‌ی k عضوی متغیرهایی مستقل با توزیع نرمال را توجیه می‌کند. این رابطه برای متد ایده آل مطرح شده در بالا ایجاد شده است، در این ایده آل دسترسی بینهایت به نمونه‌های آموزشی تابع هدف به فرض‌های قبلی اضافه شده. در عمل، با داشتن مجموعه‌ی محدود D_0 از نمونه‌های آموزشی و متد عملی جدول 5.5، این توجیه کاملاً برقرار نیست. در کل، مسئله اینجاست که تنها راه ایجاد δ_i ‌های جدید باز ترکیب D_0 با تقسیم آن به مجموعه‌های آموزشی و تست با ترکیب‌های مختلف است. بنابراین، δ_i ‌ها نیز از یکدیگر مستقل نخواهند بود، زیرا که آن‌ها از مجموعه نمونه‌های آموزشی‌ای که اشتراک دارند و از مجموعه‌ی محدود D_0 انتخاب شده‌اند (به جای اینکه با توزیع احتمال کامل \mathcal{D} انتخاب شوند).

هنگامی که تنها مجموعه‌ی محدود D_0 از نمونه‌های در دسترس است، چندین متد را می‌توان برای باز ترکیب D_0 به کاربرد. جدول 5.5 متدی به نام k -fold را که در آن مجموعه‌ی D_0 را به k زیر مجموعه‌ی هم اندازه تقسیم می‌کند. در این روش، هر نمونه‌ی D_0 دقیقاً در یک مجموعه‌ی تست استفاده و $k-1$ بار به عنوان نمونه‌ی آموزشی مورد استفاده قرار می‌گیرد. راه حل دیگر متداول انتخاب تصادفی حداقل 30 نمونه از D_0 به عنوان مجموعه‌ی تست و استفاده از بقیه‌ی نمونه‌ها برای آموزش است، این متد را می‌توان به تعداد دلخواه تکرار کرد. این متد تصادفی این مزیت را دارد که می‌توان برای کوچک کردن بازه‌های اطمینان به اندازه‌ی دلخواه، آن را بینهایت بار تکرار کرد. در مقابل، متد k -fold با تعداد داده‌های موجود با دو شرط اینکه هر نمونه تنها یک بار برای تست به کار برده می‌شود و تعداد داده‌های دسته‌ی تست حتماً باید بیشتر 30 باشند محدود می‌شود. با این وجود در متد تصادفی دیگر دسته‌های تست مستقل از همدیگر و بر اساس توزیع احتمال \mathcal{D} نخواهند

بود. در مقابل، دسته های تست ایجاد شده در روش k -fold از یکدیگر مستقل خواهند بود زیرا که هر نمونه تنها در یک دسته ی تست حضور دارد.

به طور خلاصه، هیچ فرایندی در مقایسه ی متد های یادگیری بر اساس داده های محدود تمامی ویژگی هایی که ما می خواهیم را ندارد. پس باید در نظر داشت که مدل های آماری در تست الگوریتم های یادگیری زمانی که تعداد داده های موجود محدود است به ندرت تمامی ویژگی های مورد نظر را خواهند داشت. با این وجود، این مدل ها بازه های اطمینان را که می توانند کمک بزرگی در تفسیر آزمایش های مقایسه ی متد های یادگیری است را ارائه می کنند.

5.7 خلاصه و منابع برای مطالعه ی بیشتر

نکات اصلی این فصل شامل موارد زیر می شود:

- نظریه ی آمار مبنایی برای تخمین از خطای واقعی ($error_D(h)$) فرضیه ی h بنا بر مشاهداتش از خطای مشاهده شده ($error_S(h)$) بر روی نمونه ی S ارائه می کند. برای مثال، اگر h یک فرضیه ی گسسته مقدار باشد و تعداد داده های نمونه ی S بیش از 30 باشد که به طور مستقل از یکدیگر انتخاب شده اند، آنگاه بازه ی اطمینان $N\%$ برای خطای $error_D(h)$ تقریباً بازه ی زیر خواهد بود:

$$error_S(h) \pm z_N \sqrt{\frac{error_S(h)(1 - error_S(h))}{n}}$$

در این رابطه مقدار z_N از جدول 5.1 تعیین می شود.

- در کل، مشکل تخمین بازه ی اطمینان با تعیین پارامتری که باید تخمین زده شود ($error_D(h)$) و یک تخمین زننده ($error_S(h)$) برای این کمیت انجام می گیرد. چون تخمین زننده یک متغیر تصادفی است ($error_S(h)$) وابسته به مجموعه نمونه ی تصادفی S است، آن را می توان با تابع توزیع احتمال حاکم نشان داد. بازه های اطمینان را می توان با پیدا کردن بازه ای از این تابع توزیع که $N\%$ حجم احتمال را در بر بگیرد پیدا کرد.
- یکی از دلایل خطا در دقت فرضیه تخمین زننده بایاس تخمین است. اگر Y یک تخمین زننده ی پارامتر p باشد، بایاس تخمین Y خطای بین p و مقدار امید Y خواهد بود. برای مثال اگر S داده های آموزشی برای ساخت فرضیه ی h باشد، آنگاه $error_S(h)$ نیز تخمین بایاس داری از خطای واقعی $error_D(h)$ خواهد بود.
- دلیل دوم خطا واریانس تخمین است. حتی با تخمین زننده ی بدون بایاس نیز مقدار مشاهده شده تخمین زننده در آزمایش های متفاوت با هم متفاوت است. واریانس σ^2 ی توزیع حاکم بر خواص تخمین زننده تعیین می کند که این مقدار از مقدار واقعی چقدر می تواند متفاوت باشد. این واریانس با افزایش تعداد نمونه های داده کاهش می یابد.
- مقایسه ی کارایی دو الگوریتم یادگیری نیز یک مسئله ی تخمین است، که آن زمانی که زمان و داده های آموزشی نامحدودند، بسیار ساده است اما هنگامی که منابع محدود می شوند این مسئله کمی سخت تر می گردد. یکی از راه های حل این مسئله که در این فصل

توضیح داده شده اعمال این دو الگوریتم به دو مجموعه‌ی مختلف از داده‌ها و مقایسه‌ی فرضیه‌های یاد گرفته شده با استفاده از بقیه‌ی داده‌هاست، در انتها نیز می‌توان از میانگین نتایج به عنوان اختلاف دو الگوریتم یاد کرد.

- در بسیاری موارد در نظر گرفته شده در اینجا، اشتقاق بازه‌ی اطمینان با فرض‌ها و تخمین‌هایی انجام گرفته است. برای مثال، بازه‌ی اطمینان مذکور در بالا برای $error_D(h)$ شامل تخمین توزیع دوجمله‌ای با توزیع نرمال، تخمین واریانس این توزیع و فرض اینکه توزیع احتمال حاکم بر نمونه‌ها ثابت است انجام می‌گیرد. با چنین شرایطی بازه‌ی اطمینان فقط تخمینی از بازه‌ی اطمینان خواهند بود اما با این حال آن‌ها اطلاعات مفیدی برای طراحی و بررسی نتایج یادگیری ماشین به ما می‌دهند.
- تعاریف کلیدی آماری این فصل در جدول 5.2 به طور خلاصه آورده شده است.

در بحث یافتن آماری میانگین و بررسی درستی فرضیه‌ها دریایی از اصطلاحات وجود دارد. در حالی که این فصل فقط به مفاهیم اولیه‌ی آماری می‌پردازد، می‌توانید نکات بیشتر آماری را در بسیاری از مقالات و کتب دیگر پیدا کنید. (Billingsley et al. 1986) معرفی بر مباحث آمار مربوطه ارائه می‌کند. دیگر متون آماری شامل (DeGroot 1986) و (Casella and Berger 1990) می‌شوند. (Duda and Hart 1973) نیز بررسی‌ای از این مباحث در قالب پیدا کردن عددی الگوها ارائه می‌کنند.

(Segre et al. 1991 1996)، (Etzioni and Etzioni 1994)، (Gordon and Segre 1996) بررسی‌های مهم آماری برای ارزیابی الگوریتم‌های یادگیری‌ای که کارایی‌شان با کاهش میزان محاسبات سنجیده می‌شوند ارائه می‌کنند.

(German et al. 1992) معیار مینیمم کردن بایاس و واریانس را به طور همزمان بررسی می‌کند. تحقیق بر روی بهترین راه برای یادگیری و مقایسه‌ی فرضیات بر روی تعداد محدود داده همچنان ادامه دارد. برای مثال (Dietterich 1996) مشکلات استفاده از چندین تست t جفت با استفاده از قسمت‌های مختلف داده‌های موجود به عنوان دسته‌های آموزشی و تست را بررسی می‌کند.

تمرینات

5.1 فرض کنید که فرضیه‌ای را بررسی می‌کنید که $r = 300$ خطا بر روی یک نمونه‌ی S با تعداد $n=1000$ نمونه‌ی تصادفی دارد. انحراف معیار $error_S(h)$ چقدر است؟ چگونه می‌توان این انحراف معیار را با انحراف معیار انتهای بخش 5.3.4 مقایسه کرد؟

5.2 فرضیه‌ی یاد گرفته شده‌ی h برای مفهومی منطقی را در نظر بگیرید. هنگامی که h بر روی مجموعه‌ای از 100 نمونه بررسی می‌شود 83 تای آن‌ها را درست دسته بندی می‌کند. انحراف معیار و بازه‌ی 95٪ اطمینان را برای خطای $error_D(h)$ بیابید.

5.3 فرض کنید که فرضیه‌ی h بر روی نمونه‌ای مستقل با $n=65$ دارای خطای $r=10$ است. بازه‌ی دوطرفه‌ی 90٪ اطمینان برای خطای واقعی چقدر است؟ بازه‌ی 95٪ اطمینان یک طرفه چقدر است (یعنی با احتمال 95٪ داریم $error_S(h) \leq U$)؟ بازه‌ی اطمینان 90٪ درصد یک طرفه چقدر است؟

5.4 رابطه‌ای کلی برای حد بالا و حد پایین بازه‌ی اطمینان یک طرفه‌ی N درصد برای خطاهای مختلف بین دو فرضیه با داده‌های مختلف ارائه دهید. (راهنمایی: رابطه‌ی بخش 5.5 را تغییر دهید)

5.5 توضیح دهید که چرا تخمین بازه‌ی اطمینان رابطه‌ی 5.17 به تخمین کمیت رابطه‌ی 5.16 اعمال می‌شود و چرا نمی‌توان آن را به 5.14 اعمال کرد؟

فرهنگ لغات تخصصی فصل (فارسی به انگلیسی)

paired tests	تست‌های جفت
evaluating hypotheses	ارزیابی فرضیه‌ها
confidence interval	بازه‌ی اطمینان
estimator bias	بایاس تخمین زننده
identically distributed	به طور یکسان توزیع شده‌اند
unbiased estimator	تخمین زننده‌ی بدون بایاس
sampling theory	تئوری نمونه برداری
sample error	خطای نمونه‌ای
true error	خطای واقعی
well suited	خوش تعریف
number of degrees of freedom	درجه‌ی آزادی
central limit	قضیه‌ی حد مرکزی
two sided bound	مرز دو طرفه
one side bound	مرز یک طرفه