فصل پنجم: ارزیابی فرضیه‌ها

ارزیابی تجربی دقت فرضیه‌ها اساس یادگیری ماشین است. این فصل معرفی‌ای بر متدهای تخمین دقت فرضیه‌ها با تأکید 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 به دست آوردن دقت فرضیه بر روی تمامی نمونه‌ها گمراه‌کننده 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.

۵٫۱ انگیزه

در بسیاری از موارد ارزیابی فرضیه‌ی یاد گرفته شده با حداکثر دقت ممکن بسیار مهم است. یکی از دلایل این اهمیت، تشخیص قابل‌استفاده 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 نمونه‌های آموزشی اغلب معیار ضعیفی برای دقت فرضیه بر داده‌های جدید است. زیرا که فرضیه از همین داده‌ها به وجود آمده است، پس این نمونه‌های تخمینی بایاس دار و خوش‌بینانه از دقت فرضیه بر داده‌های جدید می‌زنند. این مشکل بیشتر مواقعی پیش می‌آید که فضای فرضیه‌ای، فضایی کامل است و به فرضیه اجازه می‌دهد بر روی نمونه‌های آموزشی overfit شود. برای به دست آوردن تخمینی بدون بایاس از دقت فرضیه بر روی داده‌های جدید، معمولاً فرضیه را بر روی دسته نمونه‌های مجزایی از نمونه‌های آموزشی (دسته‌ی تست) می‌سنجیم.
- اختلاف در تخمین. دوم اینکه اگر دقت فرضیه را بر روی دسته‌ی تست که بایاس ندارند بسنجیم، با این حال امکان دارد که این دقت به دست آمده با دقت واقعی اختلاف داشته باشد، این اختلاف به چگونگی انتخاب دسته‌ی تست وابسته است. با کاهش اندازه‌ی دسته‌ی تست این میزان خطای احتمالی نیز افزایش می‌یابد.

در این فصل به متدهای ارزیابی فرضیه‌های یاد گرفته شده، متدهای مقایسه‌ی دقت دو فرضیه، و متدهای مقایسه‌ی دقت دو الگوریتم یادگیری مختلف هنگامی با داده‌ها محدودند می‌پردازیم. اکثر این مباحث بر پایه‌ی قوانین پایه‌ای آماری و تئوری نمونه‌برداری[1] هستند، البته در طول این فصل فرض شده که خواننده هیچ اطلاعات قبلی‌ای در مورد مباحث پیچیده‌ی آماری ندارد. تحقیق بر روی تست‌های آماری بررسی فرضیه‌ها خیلی وسیع است. این فصل خلاصه‌ای مقدمه‌ای از این تحقیقات آماری با تأکید بر قسمت‌هایی که بیشترین رابطه را با یادگیری و تخمین و مقایسه‌ی دقت فرضیه‌ها دارد را نیز ارائه می‌کند.

۵,۲ تخمین دقت فرضیه‌ها

زمانی که دقت یک فرضیه را تخمین می‌زنیم هدف دقت فرضیه برای دسته‌بندی نمونه‌های 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 نمادگذاری ذیل برای مسائل یادگیری استفاده خواهیم کرد. فضای نمونه‌های X (برای مثال مجموعه‌ی کل افراد جامعه) وجود دارد که تابع هدف‌های متفاوتی (مثل افرادی که می‌خواهند امسال تخته اسکی جدید بخرند) روی آن‌ها تعریف می‌شوند. ما فرض می‌کنیم که تعداد تکرار اعضای مختلف X مساوی نیست. راه‌حل ساده برای در نظر گرفتن این فرض این است که توزیع احتمال مجهول $\mathcal{D}$ را که بر روی X تعریف شده برای تعداد تکرار نمونه‌ها در نظر بگیریم (این تابع توزیع ممکن است برای افراد ۱۹ ساله خیلی بیشتر از افراد ۱۰۹ ساله باشد). توجه داشته باشید که $\mathcal{D}$ هیچ اطلاعاتی در مورد مثبت یا منفی بودن نمونه x به ما نمی‌دهد؛ این توزیع فقط احتمال برخورد با نمونه x را به ما می‌دهد. کار یادگیری، یادگیری تابع هدف f با استفاده از فضای فرضیه‌های ممکن H است. نمونه‌های آموزشی تابع هدف f توسط یک معلم به یادگیر داده می‌شود. معلم جداگانه بر اساس توزیع $\mathcal{D}$ نمونه‌ها را انتخاب می‌کند سپس نمونه‌ی x را با مقدار تابع هدف f(x) به یادگیر می‌دهد.

برای تصور، تابع هدف "افرادی که می‌خواهند امسال تخته اسکی جدید بخرند" را با نمونه‌های آموزشی‌ای که از تحقیقی که از افرادی که به پیست اسکی وارد می‌شوند به دست آمده در نظر بگیرید. در این مثال، فضای نمونه‌های X کل افراد جامعه است، این نمونه‌ها با ویژگی‌هایی نظیر سن، شغل، تعداد دفعات اسکی در سال و غیره توصیف می‌شوند. توزیع $\mathcal{D}$ برای هر فرد x احتمال اینکه فرد بعدی‌ای باشد که وارد پیست می‌شود را می‌دهد. تابع هدف $f:X\rightarrow\{0,1\}$ افراد را بر اساس اینکه قصد خرید تخته اسکی جدید دارند دسته‌بندی می‌کند.

[1] sampling theory

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

۱. 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؟
۲. خطای احتمالی این تقریب دقت چقدر است؟

۵,۲,۱ 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 نمونه‌ای[۲] و خطای واقعی[۳] می‌نامیم.

خطای نمونه‌ای یک فرضیه بر اساس مجموعه نمونه‌های S از X نسبتی از S است که فرضیه اشتباه دسته‌بندی 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 δ(f(x),h(x)) اگر $f(x) \neq h(x)$ یک است و در غیر این صورت صفر می باشد.

خطای واقعی یک فرضیه احتمال این است که فرضیه نمونه‌ای که با توزیع $\mathcal{D}$ انتخاب شده را اشتباه دسته‌بندی 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 نمونه‌ی انتخابی با توزیع $\mathcal{D}$ را اشتباه دسته‌بندی 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 ی که با توزیع $\mathcal{D}$ انتخاب شده باشد نشان می دهد.

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

---

[۲] sample error
[۳] true error

۵,۲,۲ بازه‌های اطمینان برای فرضیه‌های گسسته مقدار

چگونه به سؤال "$error_S(h)$ با چه میزان دقت $error_{\mathcal{D}}(h)$ 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 نمونه‌های S در شرایط زیر تعیین کنیم:

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

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

۱. بدون داشتن اطلاعات بیشتر، محتمل‌ترین مقدار $error_{\mathcal{D}}(h)$، $error_S(h)$ است.
۲. با احتمال تقریباً ۹۵٪ خطای واقعی $error_{\mathcal{D}}(h)$ در بازه‌ی زیر است:

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

برای تصور، مجموعه‌ی داده‌های 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 نمونه‌ای $error_S(h)$=12/40=0.3 است. بدون داشتن اطلاعات بیشتر، بهترین تخمین برای $error_{\mathcal{D}}(h)$ همان مقدار ۰,۳ است. با این وجود، انتظار نمی‌رود که این تخمین، تخمین کاملی از خطای واقعی باشد. اگر دسته‌ی دیگری از نمونه‌ها مثل S′ که ۴۰ نمونه دارد داشته باشیم، قاعدتاً انتظار داریم که $error_{S'}(h)$ تقریباً با $error_S(h)$ مساوی باشد. اختلاف احتمالی این دو مقدار به چینش دو مجموعه‌ی S و S′ وابسته است. در واقع اگر این آزمایش را بارها تکرار کنیم و در هر بار تکرار از مجموعه‌ی ۴۰ نمونه‌ای S_i استفاده کنیم، به این نتیجه خواهیم رسید که تقریباً در ۹۵٪ این آزمایشات بازه‌ی محاسبه شده خطای واقعی را شامل می‌شود. به همین دلیل به این بازه، بازه‌ی اطمینان ۹۵ درصدی تخمین خطای واقعی $error_{\mathcal{D}}(h)$ می‌گویند. در مثال فعلی که در آن r = 12 و n = 40 است، بازه‌ی ۹۵٪ بر اساس داده‌های بالا 0.30±(1.96·.07)=0.30±.14 خواهد بود.

رابطه‌ای که در بالا برای بازه‌ی اطمینان ۹۵ درصدی بیان شد را می‌توان برای اطمینان N درصدی تعمیم داد. مقدار ثابت 1.96 که در تعریف رابطه آمده برای اطمینان ۹۵ درصدی است، با تغییر دادن این ثابت، z_N، می‌توان بازه‌ی اطمینان N% را محاسبه کرد. رابطه‌ی کلی محاسبه‌ی بازه‌ی اطمینان N% برای خطای $error_{\mathcal{D}}(h)$ در زیر آمده است:

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

در این رابطه ثابت z_N بر حسب درصد اطمینان تغییر می‌کند. مقادیر نظیر بعضی درصدها در جدول ۵,۱ آمده است.

خطای N درصدی:	۵۰٪	۶۸٪	۸۰٪	۹۰٪	۹۵٪	۹۸٪	۹۹٪

ثابت z_N:	۰٫۶۷	۱٫۰۰	۱٫۲۸	۱٫۶۴	۱٫۹۶	۲٫۳۳	۲٫۵۸

جدول ۵٫۱ مقادیر $\boldsymbol{z_N}$ 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 می یابد.

رابطه‌ی ۵٫۱ نحوه‌ی محاسبه‌ی بازه‌های اطمینان یا ستون‌های خطا[۴] 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 مستقل‌اند. 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 ۳۰ است و $error_S(h)$ خیلی نزدیک ۰ یا ۱ نیست دقت خوبی دارد. به عبارت دقیق‌تر این رابطه زمانی که نامساوی زیر صادق است دقت خوبی دارد:

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

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

۵٫۳ اساس تئوری نمونه‌برداری

در این قسمت ایده‌های اصلی آماری و تئوری نمونه‌برداری را معرفی خواهیم کرد، این ایده‌ها شامل توزیع‌های احتمال، امید ریاضی، واریانس، توزیع‌های دوجمله‌ای 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. خوانندگانی 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 ۵٫۲ 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 می کند.
- 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 Var(Y)=$E[(Y - \mu_Y)^2]$ تعریف می‌شود. واریانس پهنای توزیع را حول میانگین بیان

[۴] 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 دوجمله‌ای 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 بازه این است که با احتمال N%، p را در بر خواهد گرفت.

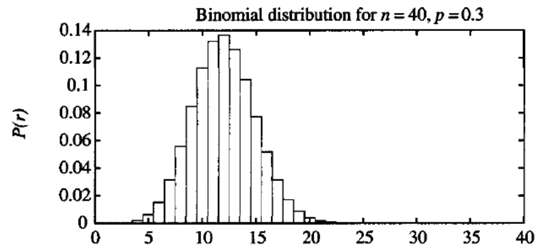
جدول ۵,۲ تعاریف و حقایق پایه‌ای آمار

۵,۳,۱ تخمین خطا و تخمین ویژگی‌های توزیع دوجمله‌ای

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 دسته‌بندی شدن توسط فرضیه‌ی h است.

جواب این سؤال در توجه به این حقیقت است که اندازه‌گیری 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 اندازه‌گیری متغیرهای تصادفی $error_{S_1}(h)$، $error_{S_2}(h)$، ....، $error_k(h)$ 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 ۵,۳ نزدیک خواهد شد. این نمودار توزیع احتمال خاصی به نام توزیع دوجمله‌ای را نشان می‌دهد.



توزیع احتمال دوجمله‌ای برای r بار شیر آمدن در آزمایشی که n سکه‌ی مجزا پرتاب می‌شوند، احتمـال شـیر آمـدن در هـر کـدام از سکه‌ها p است. تابع توزیع به شکل زیر تعریف می‌شود:

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

اگر متغیر تصادفی X از توزیع دوجمله‌ای پیروی کند:

- احتمال اینکه Pr(X=r) (X مقدار r را بگیرد) با P(r) مشخص می‌شود.
- امید X ، یا همان میانگین X، E[X] از رابطه‌ی زیر به دست خواهد آمد:

  E[X]=np

- واریانس X، Var(X) از رابطه‌ی زیر به دست خواهد آمد:

  Var(X)=np(1-p)

- انحراف معیار X، σ_X از رابطه‌ی زیر به دست خواهد آمد:

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

برای nهای به اندازه‌ی کافی بزرگ توزیع دوجمله‌ای تقریباً نزدیک به توزیع نرمال با همان واریانس و میانگین خواهـد بـود (جـدول ۵,۴). توصیه می‌شود که فقط زمانی که $np(1-p) \geq 5$ باشد از این تقریب استفاده کرد.

جدول ۵,۳ توزیع دوجمله‌ای

۵,۳,۲ توزیع دوجمله‌ای

یکی از روش‌های خوب درک یادگیری توزیع دوجمله‌ای بررسی مسئله‌ی ذیل است. سکه‌ای معیوب (کج) به شما داده می‌شود و از شما خواسـته می‌شود تا احتمال اینکه سکه پس از پرتاب شیر بیاید را حساب کنید. بیایید احتمال شیر آمدن سکه‌ی معیوب را با p نشان دهیم. شما سکه را n بار پرتاب می‌کنید، از این n بار r بار شیر می‌آید. یک تخمین منطقی از p مقدار r/n است. توجه دارید که اگر این آزمایش را تکرار کنـیم و n بار سکه را پرتاب کنیم، انتظار نمی‌رود که تعداد شیرهای آمده دقیقاً r قبلی باشد، پس بنابراین مقدار p به دست آمده نیز با مقدار p قبلی یکـی نخواهد بود. توزیع دوجمله‌ای احتمال وقوع مقدار r (بین ۰ تا n) را در n پرتاب مشخص می‌کند، این احتمال با فـرض اینکـه تمـامی پرتاب‌هـا مستقل و احتمال شیر آمدن دقیقاً p است محاسبه می‌شود.

جالب است که بدانیم، تخمین p از یک دسته پرتاب مشابه تخمین $error_{\mathcal{D}}(h)$ بر اساس یک دسته نمونه است. شیر بودن یک پرتاب مشابه اشتباه دسته‌بندی شدن یک نمونه‌ی تصادفی (با توزیع $\mathcal{D}$) است. احتمال p یا همان احتمال شیر آمدن یک پرتاب مشابه احتمال اشتباه دسته‌بندی شدن یک نمونه تصادفی (یا همان $error_{\mathcal{D}}(h)$) است. تعداد r یا همان تعداد شیرها در n پرتاب نیز مشابه تعداد دسته‌بندی‌های اشتباه نمونه‌ها از n نمونه‌ی تصادفی است. بنابراین r/n مشابه $error_S(h)$ است و مسئله‌ی تخمین p در سکه نیز مشابه مسئله‌ی تخمین $error_{\mathcal{D}}(h)$ در فرضیه‌ها است. توزیع دوجمله‌ای فرم کلی توزیع احتمال را برای متغیر تصادفی r مشخص می‌کند حال فرقی نمی‌کند که این r تعداد شیرها باشد یا تعداد دسته‌بندی‌های اشتباه. فرم دقیق‌تر توزیع دوجمله‌ای به تعداد نمونه‌ها و p (یا همان $error_{\mathcal{D}}(h)$) وابسته است.

شرایطی که توزیع دوجمله‌ای در آن صادق است:

۱. مبنای کار یک آزمایش (مثل پرتاب سکه) است که خروجی‌اش به عنوان متغیر تصادفی (مثل Y) است. مقدار تصادفی Y می‌تواند فقط دو مقدار داشته باشد (برای مثال Y=1 برای شیر و Y=0 برای خط)

۲. احتمال اینکه Y=1 شود در هر آزمایش مبنا 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.

۳. سری‌ای از آزمایش‌های مبنا پشت سر هم انجام می‌شود (مثل پرتاب‌های سکه) و سری‌ای از متغیرهای تصادفی هم ارزش مثل Y_1، Y_2، ...،Y_n را ایجاد می‌کند. اگر R تعداد آزمایش‌ها که در آن‌ها $Y_i=1$ است در نظر بگیریم خواهیم داشت:

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

۴. احتمال اینکه متغیر تصادفی R مقدار r باشد (احتمال اینکه دقیقاً r بار شیر بیاید) بر اساس توزیع دوجمله‌ای به صورت زیر است:

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

نموداری از این رابطه در جدول ۵,۳ آمده بود.

توزیع دوجمله‌ای احتمال تعداد خطای r در میان n نمونه را مشابه احتمال r بار شیر آمدن در میان n آزمایش پرتاب سکه توصیف می‌کند.

۵,۳,۳ میانگین و واریانس

دو خاصیتی که معمولاً در مورد متغیرهای تصادفی مطرح می‌شود امید (مقدار انتظاری یا میانگین) و واریانس است. مقدار امید، میانگین مقادیر تصادفی به دست آمده بعد از آزمایش‌های بسیار است. به عبارت دقیق‌تر:

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

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

برای مثال اگر متغیر تصادفی Y مقدار ۱ را با احتمال ۰,۷ و مقدار ۲ را با احتمال ۰,۳ بپذیرد مقدار امید (1·0.7+2·0.3=1.3) خواهد بـود. در توزیع دوجمله‌ای این تعریف به شکل زیر تغییر شکل پیدا می‌کند:

$$E[Y] = np \tag{5.4}$$

در این رابطه n و p پارامترهای توزیع دوجمله‌ای در رابطه‌ی ۵,۲ هستند.

کمیت دوم مطرح "پهنا[۵]" یا "میزان پخشی[۶]" 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:

$$Var[Y] \equiv E[(Y - E[Y])^2] \tag{5.5}$$

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

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

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

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

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

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

۵,۳,٤ تخمین زننده‌ها، بایاس و واریانس

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

بیایید $error_S(h)$ و $error_D(h)$ را با استفاده از رابطه‌ی ۵,۲ که توزیع دوجمله‌ای را بیان می‌کند توصیف کنیم. داریم که

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

$$error_D(h) = p$$

[۵] width
[۶] spread

در این رابطه n تعداد نمونه‌های مجموعه‌ی S و r تعداد دسته‌بندی‌های اشتباه h از مجموعه‌ی S است و p نیز احتمال دسته‌بندی اشـتباه h از نمونه‌ای انتخاب شده با توزیع $\mathcal{D}$ است.

متخصصان $error_S(h)$ را تخمین زننده‌ای[۷] از خطای واقعی $error_D(h)$ می‌نامند. در کل، تخمین زننده‌ی یـک مقـدار تصـادفی بـرای تخمین ویژگی‌های جمعیت آن متغیر تصادفی به کار می‌رود. اولین سؤالی که درباره‌ی هر تخمین زننده مطرح می‌شود این است که آیـا تخمـین زننده در میانگین تخمین درستی به ما می‌دهد؟ بایاس تخمین را به عنوان اختلاف بین مقدار امید تخمین زننده و مقـدار واقعـی متغیـر تصـادفی تعریف می‌کنیم.

تعریف: بایاس تخمین[۸] برای تخمین زننده‌ی Y از پارامتر p به صورت

$$E[Y]-p$$

تعریف می‌شود.

اگر مقدار بایاس تخمین زننده صفر باشد می‌گوییم که Y یک تخمین زننده‌ی بدون بایاس از p است. توجه داشته این حالتی اسـت کـه پـس از تعداد زیادی آزمایش تصادفی میانگین مقدار تصادفی به امید تخمین زننده میل کند.

آیا $error_S(h)$ تخمین زننده‌ی بدون بایاسی از $error_D(h)$ است؟ بله، زیرا که مقدار امید r در توزیـع دوجملـه‌ای np اسـت (رابطـه‌ی ۵,۴). حال چون که n ثابت است، پس مقدار امید r/n همان p است.

دو نکته‌ی قابل‌توجه در بایاس تخمینی وجود دارد. اول، همان طور که در ابتدای این فصل نیز گفته شد، بررسـی فرضـیه‌ها بـر روی نمونـه‌های آموزشی، تخمینی بایاس دار از خطای فرضیه به ما می‌دهد، این دقیقاً همان نکته‌ای است که بایاس تخمین به آن اشاره مـی‌کنـد. بـرای اینکـه $error_S(h)$ تخمینی بدون بایاس از $error_D(h)$ به ما بدهد، باید فرضیه‌ی h و نمونه‌های S باید مستقل باشند. دوم اینکه ایـن مفهـوم نباید با بایاس استقرایی که در فصل ۲ بیان شد اشتباه گرفته شود. بایاس تخمینی یک مقدار عددی است در حالی که بایاس استقرایی دسـته‌ای از پیش‌فرض‌ها[۹] است.

ویژگی مهم دیگر هر تخمین زننده مقدار واریانس آن است. با داشتن انتخاب بین تخمین زننده‌های بدون بایاس مختلف، قابـل‌درک اسـت کـه تخمین زننده‌ای را انتخاب کنیم که کمترین مقدار واریانس را داشته باشد. با تعریفی که از واریانس ارائه شد، این انتخـاب باعـث مـی‌شـود کـه خطای انتظاری بین تخمین و مقدار واقعی به کمترین مقدار برسد.

برای تصور این مفاهیم، فرض کنید که می‌خواهیم فرضیه‌ای را بررسی کنیم که r=12 خطـا بـر روی نمونـه‌هایی بـا تعـداد n=40 دارد. اگـر $error_S(h)$ یک تخمین زننده‌ی 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 توزیـع دوجملـه‌ای انتخـاب مـی‌شـود بـرای واریـانس از رابطه‌ی 5.7 داریم: np(1-p). متأسفانه هنوز مقدار p مجهول است، اما می‌توان بجـای آن از تخمـین r/n مـان از p اسـتفاده کنـیم. پـس

---

[۷] estimator

[۸] estimation bias

[۹] assertion

واریانس r خواهد بود 8.4=(1-0.3)40·0.3 پس مقدار انحراف معیار $\sqrt{8.4} \approx 2.9$. پس انحراف معیار r/n، 2.9/40=.07 خواهد بود. به طور خلاصه، $error_S(h)$ در این مثال مقدار امید ۰٫۳۰ با انحراف معیار تقریبا ۰٫۰۷ است (تمرین ۵٫۱).

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

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

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

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

۵٫۳٫۵ بازه‌ی اطمینان

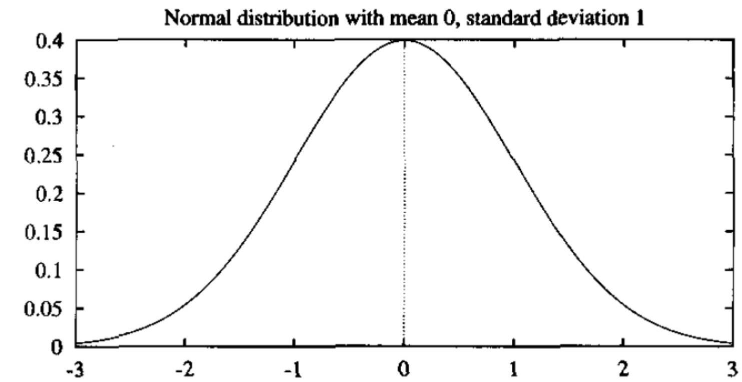
یکی از راه‌های معمول توصیف عدم قطعیت یک تخمین توجه به بازه و احتمالی است که انتظار می‌رود که مقدار واقعی در این بازه باشد است. چنین تخمینی، تخمین بازه‌ی اطمینان نامیده می‌شود.

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 ۹۵٪ مقدار $error_D(h)$ در بازه‌ی 0.30±0.14 است.

بازه‌های اطمینان $error_D(h)$ چگونه به دست می‌آیند؟ جواب در این حقیقت نهفته است که توزیع احتمال دوجمله‌ای بر $error_S(h)$ حاکم است. میانگین این توزیع مقدار $error_D(h)$ است و انحراف معیار نیز از رابطه‌ی ۵٫۹ به دست می‌آید. بنابراین، برای به دست آوردن بازه‌ی ۹۵٪ فقط نیاز است که بازه را حول مقدار میانگین $error_D(h)$ بگیریم تا ۹۵٪ از کل احتمال را در بر بگیرد. این بازه، بازه‌ای حول $error_D(h)$ که در 95% موارد $error_S(h)$ درون آن قرار می گیرد.

برای عدد معلوم N چگونه می‌توان اندازه‌ی بازه‌ی N% احتمال را به دست آورد؟ متأسفانه، این محاسبه برای توزیع احتمال دوجمله‌ای زمان‌بر است. خوشبختانه، با وجود زمان‌بری، در اکثر موارد تقریب خوبی برای بازه به دست می‌آید، زیرا که با تعداد نسبتاً زیاد نمونه توزیع به توزیع نرمال میل می‌کند. توزیع نرمال (که در جدول ۵٫۴ نیز آمده) شاید خوش تعریف‌ترین توزیع احتمال باشد. همان طور که در جدول ۵٫۴ 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 دوجمله‌ای 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 به دست خواهد آمد:

$$\int_a^b \mathrm{p(x)dx}$$

- امید یا میانگین X، E[X]:

$$E[X] = \mu$$

- واریانس X، Var(X):

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

- انحراف معیار X، σ_X

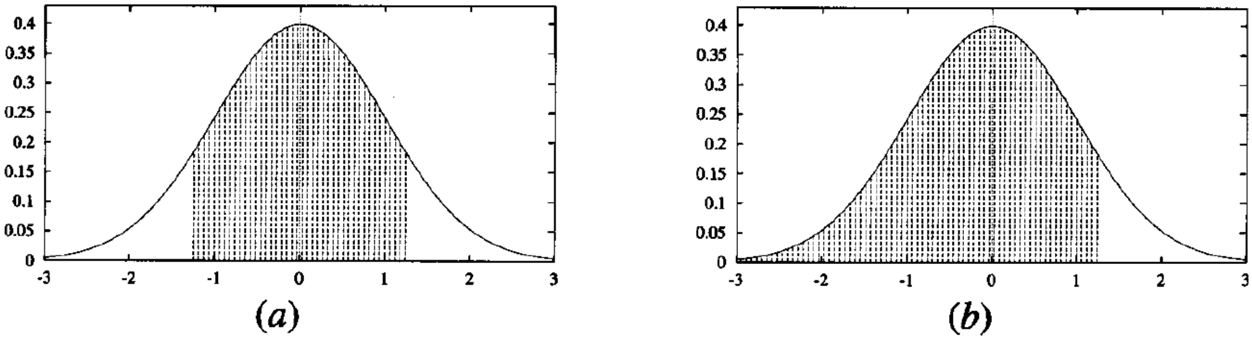
$$\sigma_X = \sigma$$

قضیه‌ی حد مرکزی[۱] (در بخش ۵,۴,۱) نشان می‌دهد که مجموع متغیرهای تصادفی را با توزیع دلخواه می‌توان با توزیع نرمال بررسی کرد.

جدول ۵,۴ 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 ۵,۱ نیز از این حقیقت به دست آمده. ثابت z_N که در جدول ۵,۱ 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.

[۱] central limit



شکل ۵,۱ توزیع نرمال با میانگین ۰ و انحراف معیار ۱.

(a) با احتمال ۸۰٪ 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 ۱۰٪ متغیر تصادفی در سمت راست و با احتمال ۱۰٪ متغیر تصادفی در سمت چپ این بازه قرار می‌گیرد. (b) با احتمال ۹۰٪ متغیر تصادفی در بازه‌ی از یک طرف محدود (-∞,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:

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

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

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

این واقعیت را می‌توان به سادگی با واقعیت‌های کلی قبلی ذکر شده در مورد بازه‌ی N% در توابع گسسته مقدار ترکیب کرد (رابطه‌ی ۵,۱). ابتدا اینکه می‌دانیم $error_S(h)$ از توزیع دوجمله‌ای پیروی می‌کند که میانگین آن $error_D(h)$ است و انحراف معیارش نیز از رابطه‌ی ۵,۹ به دست می‌آید. دوم اینکه می‌دانیم که برای زمانی که تعداد n به اندازه‌ی کافی بزرگ باشد توزیع دوجمله‌ای را می‌توان با تقریب خوبی با توزیع نرمال تقریب زد. سوم اینکه رابطه‌ی ۵,۱۱ 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 ۵,۱۱ برای توابع گسسته مقدار به رابطه‌ی ۵,۱ می‌رسیم.

$$error_S(h) \pm z_N \sqrt{\frac{error_S(h)\left(1 - error_{S(h)}\right)}{n}}$$

دو نکته‌ی مهم در تخمین این رابطه به شرح زیرند:

۱. در تخمین انحراف معیار σ برای $error_S(h)$، ما از $error_S(h)$ به جای $error_D(h)$ استفاده کردیم (در نتیجه‌گیری رابطه‌ی ۵,۹ از رابطه‌ی ۵,۸)
۲. توزیع دوجمله‌ای را با توزیع نرمال تخمین زده‌ایم.

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

۵,۳,۶ بازه‌های یک طرفه و دو طرفه

توجه دارید که بازه‌ی اطمینان ذکر شده مرز دوطرفه[۲] 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 طرفه[۳]). برای مثال ممکن است جواب سؤال "احتمال اینکه $error_{\mathcal{D}}(h)$ 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 آن را 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 بازه‌ای با مرز یک طرفه تبدیل کرد (مثل شکل (b) 5.1). اگر بازه‌ی اولیه اطمینان 100(1-α)% داشته باشد، بازه‌های یک طرفه‌ی فقط محدود از بالا اطمینان 100(1-α/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 فرضیه‌ای با r=12 خطا بر روی مجموعه‌ای با n=40 نمونه داریم که نمونه‌ها از فرضیه مستقل‌اند. همان طور که بالاتر نیز گفته شد، این اطلاعات بازه‌ی 95% (دوطرفه) ی 0.30±0.14 را مشخص می‌کند. در اینجا، 100(1-α)=95% پس α=0.05. بنابراین بدون اضافه کردن هیچ پیش‌فرض اضافه‌ای، می‌توانیم بگوییم که با اطمینان 100(1-α/2)=97.5%، $error_{\mathcal{D}}(h)$ حداکثر 0.30+0.14=.44 است. بنابراین، مرزی یک طرفه برای $error_{\mathcal{D}}(h)$ با اطمینان دو برابر نسبت به مرز دوطرفه خواهیم داشت (تمرین ۵,۳).

۵,۴ روش کلی برای استخراج بازه‌های اطمینان

در قسمت قبل چگونگی به دست آوردن بازه‌های تخمین برای حالت خاص: تخمین $error_{\mathcal{D}}(h)$ برای توابع گسسته مقدار h بر پایه‌ی مجموعه‌ای با 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 زیرمجموعه‌ای با n عضو دانست. فرایند کلی مراحل زیر را در بر می‌گیرد:

۱. معلوم کردن پارامتر p ای که از مجموعه می‌خواهیم تخمین بزنیم، برای مثال $error_{\mathcal{D}}(h)$.
۲. تعریف تخمین زننده‌ی Y، مثل ، مثل $error_{S}(h)$. بهتر است این تخمین زننده واریانس کم داشته و بدون بایاس باشد.
۳. مشخص کردن توزیع احتمال $\mathcal{D}_Y$ که کار تخمین زننده‌ی Y را کنترل می‌کند. مشخص کردن این توزیع شامل مشخص کردن واریانس و میانگین نیز می‌شود.

---

[۲] two sided bound

[۳] one side bound

۴. مشخص کردن بازه‌ی N% با پیدا کردن مقادیر آستانه‌ی L و U که N% از جرم احتمال توزیع $\mathcal{D}_Y$ 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 بیایید نتیجه‌ی اساسی قضیه‌ی حد مرکزی را بررسی کنیم.

۵,٤,۱ قضیه‌ی حد مرکزی

یکی از حقیقت‌هایی که برای پیدا کردن بازه‌های 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 توزیع[۴] 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→∞ 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 ۵,۱ قضیه‌ی حد مرکزی. فرض کنید که دسته‌ای 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→∞ توزیع احتمال حاکم بر

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

به توزیع نرمال با میانگین صفر و انحراف معیار ۱ 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 می کند.

قضیه‌ی حد مرکزی، حقیقتی پرکاربرد است، زیرا که به هر صورت که یک تخمین زننده برای تخمین میانگین چیزی تعریف کنیم ($error_S(h)$ تخمین زننده‌ی میانگین خطاست)، برای n های به اندازه‌ی کافی بزرگ آن را می‌توان با توزیع نرمال تخمین زد. حال اگر واریانس این توزیع نرمال (تخمینی) را بدانیم می‌توانیم با استفاده از رابطه‌ی ۵,۱۱ برای محاسبه‌ی بازه‌های 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 $error_S(h)$ را توصیف می‌کرد استفاده کردیم.

---

[۴] identically distributed

۵,۵ تفاوت خطاهای دو فرضیه

حالتی را تصور کنید که دو فرضیه‌ی h_2 و h_1 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.

$$d \equiv error_{\mathcal{D}}(h_1) - error_{\mathcal{D}}(h_2)$$

در اینجا ما از فرایند چهار مرحله‌ای ارائه شده در ابتدای بخش ۵,۴ برای به دست 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:

$$\hat{d} \equiv error_{\mathcal{D}}(h_1) - error_{\mathcal{D}}(h_2)$$

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

اما $\hat{d}$ از چه توزیع احتمالی پیروی می‌کند؟ با استفاده از آنچه در قسمت‌های قبلی گفته شد، اگر n_2 و n_1 به اندازه‌ی کافی بزرگ باشند (هر دو بزرگ‌تر از ۳۰ باشند) هر دو متغیر تصادفی $error_{S_2}(h_2)$ و $error_{S_1}(h_1)$ 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 $error_{S_2}(h_2)$ و $error_{S_1}(h_1)$ است. با استفاده از رابطه‌ی 5.9 برای تخمین واریانس هر یک از این دو توزیع داریم که:

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

حال که توزیع احتمالی که $\hat{d}$ را کنترل می‌کند را مشخص کرده‌ایم، به راحتی می‌توان بازه‌های اطمینان را که از $\hat{d}$ برای d به دست 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:

$$\hat{d} \pm z_N \sqrt{\frac{error_{S_1}(h_1)\left(1 - error_{S_1}(h_1)\right)}{n_1} + \frac{error_{S_2}(h_2)\left(1 - error_{S_2}(h_2)\right)}{n_2}} \quad (5.13)$$

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 ۵,۳,۶ به بازه‌های 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 ۵٫۱۳) 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) قابل‌قبول 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 رابطه‌ی ۵٫۱۲ است، زیرا که دو مجموعه‌ی S_1 و S_2 هر دو S در نظر گرفته شده‌اند. دلیل این کاهش استفاده از یک مجموعه‌ی نمونه‌ای S برای ارزیابی اختلاف اثر اختلافات تصادفی بین ترکیب S_1 و S_2 را حذف خواهد کرد. در این حالت، بازه‌ی اطمینان رابطه‌ی ۵٫۱۳ در کل بازه‌ای محافظه کارانه، اما هنوز درست، خواهد بود.

۵٫۵٫۱ تست فرضیه[5]

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 $error_{\mathcal{D}}(h_1) > error_{\mathcal{D}}(h_2)$ ؟" 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 ۱۰۰ 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 $error_{\mathcal{D}}(h_1) \leq error_{\mathcal{D}}(h_2)$ است نیز مشاهده شود. احتمال اینکه داشته باشیم $error_{\mathcal{D}}(h_1) > error_{\mathcal{D}}(h_2)$ 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 $\hat{d} < d + .10$ 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 $\hat{d} < \mu_{\hat{d}} + .10$ 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 اندازه‌گیری خواهد شد.

بیایید این محاسبه را با بازنویسی دوباره‌ی $\hat{d} < \mu_{\hat{d}} + .10$ با انحراف از معیار شروع کنیم. با استفاده از رابطه‌ی ۵٫۱۲ می‌توان به دست 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 ۵٫۱، می‌توان به دست آورد که بازه‌ای تا ۱٫۶۴ برابر انحراف حول میانگین برای بازه‌ی دوطرفه احتمال ۹۰٪ دارد. بنابراین، بازه‌ی اطمینان دو طرفه احتمال ۹۵٪ را خواهد داشت.

---

[5] Hypothesis testing

بنابراین، با داشتن مشاهده‌ی $\hat{d} = .10$، احتمال اینکه $error_{\mathcal{D}}(h_1) > error_{\mathcal{D}}(h_2)$ تقریباً $.95$ است. در واژگـان ادبیـات آمـاری، می‌گوییم که این فرضیه که "$error_{\mathcal{D}}(h_1) > error_{\mathcal{D}}(h_2)$" را با اطمینان $.95$ می‌پذیریم. یا به طور مشابه ممکن است بگوییم که فرضیه متضاد[۶] (یا فرضیه‌ی تهی[۷]) را با احتمال 05.=(95.-1) رد می‌کنیم.

۵٫۶ مقایسه‌ی الگوریتم‌های یادگیری

بعضی مواقع مقایسه‌ی عملکرد دو الگوریتم یادگیری L_A و L_B برای ما از مقایسـه‌ی دو فرضـیه اهمیـت بیشـتری دارد. آزمـون مناسـب بـرای مقایسه‌ی الگوریتم‌های یادگیری چیست و چگونه می‌توان معلوم کرد که تفاوت‌های به دست آمده از نظر آماری قابل‌توجه‌انـد؟ بـا وجـود اینکـه بحث‌ها هنوز در این مبحث از یادگیری ماشین داغ است اما ما در اینجا روشی خاص را معرفـی خـواهیم کـرد. بحـث در مـورد دیگـر متـدهای 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 الگـوریتم L_A یا L_B به طور متوسط برای یادگیری تابع هدف f معلوم متناسب‌ترند. یکی از راه‌های تعریف "به طور متوسـط" بررسـی کـارایی نسـبی دو الگوریتم بر روی تمامی مجموعه نمونه‌های n عضوی ممکن که با استفاده از توزیع احتمال نمونه‌ای $\mathcal{D}$ است. به عبارت دیگر، قصد داریـم کـه مقدار امید خطای بین دو فرضیه را تخمین بزنیم

$$\underset{S\subset\mathcal{D}}{\mathrm{E}}\left[error_{\mathcal{D}}\big(L_A(S)\big) - error_{\mathcal{D}}\big(L_B(S)\big)\right] \qquad (5.14)$$

در این رابطه L(S) فرضیه‌ای است که با استفاده از متد L از نمونه‌های S به دست می‌آید، $S\subset\mathcal{D}$ نیز به این معناست کـه مقـدار امیـد بـر روی نمونه‌های S که با توزیع $\mathcal{D}$ انتخاب می‌شوند محاسبه می‌شود. عبارت بالا مقدار امید اختلاف بین خطاهای یادگیری دو متـد L_A و L_B را نشـان می دهد.

البته در عمل مجموعه‌ای محدود از نمونه‌ها D_0 در دسترس است و بررسی بین دو متد را بر روی این مجموعه‌ی محدود انجـام مـی‌دهـیم. 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 آموزشـی 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 فرضـیه‌های یـاد گرفتـه شده استفاده کرد. به عبارت دیگر، ما کمیت زیر را محاسبه می‌کنیم:

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

توجه داشته باشید که این تخمین زننده و کمیت رابطه‌ی ۵٫۱۴ دو تفاوت کلیدی دارند. ابتدا اینکه در این رابطه از $error_{T_0}(h)$ برای تخمین مقدار $error_{\mathcal{D}}(h)$ استفاده شده است. دوم اینکه در این رابطه فقط تفاوت بین خطاها بـرای یـک مجموعـه‌ی آموزشـی S_0 بـه جـای کـل مجموعه‌های آموزشی ممکن توزیع $\mathcal{D}$ استفاده شده است.

---

[۶] opposite hyposesis
[۷] null hyposesis

یکی از راه‌های بهبود این تخمین زننده‌ی رابطه‌ی ۵٫۱۵ تقسیم داده‌های در چندین مرحله D_0 به مجموعه‌های کوچک‌تر و استفاده از میانگین خطای به دست آمده از دسته‌ی تست در آزمایش‌های مختلف است. این کار به فرایند نشان داده شده در جدول ۵٫۵ برای مقایسه‌ی خطاهای دو متد یادگیری بر اساس داده‌های ثابت D_0 می‌انجامد. این فرایند تقسیم ابتدا داده‌ها را به k زیرمجموعه‌ی هم‌اندازه که هر کدام حداقل ۳۰ نمونه دارند تقسیم می‌کند. سپس الگوریتم یادگیری را k بار آموزش می‌دهد و آزمایش می‌کند، در هر یک از این k بار یکی از k زیرمجموعه به عنوان مجموعه‌ی تست مورد استفاده قرار می‌گیرد و بقیه داده‌ها نیز مجموعه‌ی آموزشی خواهند بود. به این ترتیب، الگوریتم‌های یادگیری بر روی k مجموعه‌ی مجزای تست بررسی می‌شوند و میانگین اختلاف در خطاهای $\bar{\delta}$ به عنوان یک تخمین زننده برای اختلاف بین دو الگوریتم یادگیری انتخاب می‌شود.

کمیت $\bar{\delta}$ که از فرایند جدول ۵٫۵ به دست می‌آید را می‌توان به عنوان تخمینی از کمیت مطلوب رابطه‌ی ۵٫۱۴ به حساب آورد. حتی می‌توان به $\bar{\delta}$ به دید تخمینی از کمیت زیر نگاه کرد:

$$\underset{s\subset D_0}{\mathrm{E}}[error_{\mathcal{D}}\left(L_A(S)\right) - error_{\mathcal{D}}(L_B(S))] \tag{5.16}$$

در این رابطه S مجموعه‌ای دلخواه از $\frac{k-1}{k}|D_0|$ نمونه است که با توزیع یکنواخت از D_0 انتخاب شده‌اند. تنها تفاوت بین این کمیت و کمیت اصلی ما در رابطه‌ی ۵٫۱۴ این است که این کمیت مقدار امید را بر روی زیرمجموعه‌ای از داده‌های موجود D_0 پیدا می‌کند به جای اینکه از تمامی نمونه‌ها با توزیع $\mathcal{D}$ استفاده کند.

---

۱. داده‌های موجود D_0 را به k دسته‌ی مجزای $T_1, T_2, \dots, T_k$ با اندازه‌های مساوی تقسیم کن، اندازه‌ی هر مجموعه باید حداقل ۳۰ باشد.

۲. برای تمامی مقادیر i با شروع از ۱ و کمتر مساوی از k:

از T_i برای دسته‌ی تست و از بقیه‌ی داده‌ها برای دسته‌ی آموزشی S_i استفاده کن.

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

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

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

جدول ۵٫۵ فرایند تخمین تفاوت بین خطاهای بین دو متد یادگیری $\boldsymbol{L_A}$ و $\boldsymbol{L_B}$. بازه‌های اطمینان این تخمین در متن آورده شده اند.

بازه‌ی اطمینان N% برای تخمین کمیت رابطه‌ی ۵٫۱۶ با $\bar{\delta}$ به فرم زیر است:

$$\bar{\delta} \pm t_{N,k-1}s_{\bar{\delta}} \tag{5.17}$$

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

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

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 آزادی[۸] نیز نامیده می‌شود و معمولاً با υ 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 ۵،۶ آورده شده، توجه دارید که با k→∞ مقدار $t_{N,k-1}$ به ثابت z_N میل می کند.

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 ۵،۵ در مورد مقایسه‌ی فرضیه‌ها با دسته تست‌های مستقل گفته شد در تضاد است. به تست‌هایی که فرضیه‌ها بر روی مجموعه‌های مشابهی تست می‌شوند تست‌های جفت[۹] می‌گویند. تست‌های جفت معمولاً بازه‌های اطمینان کوچک‌تری ایجاد می‌کنند زیرا که در خطاهای مشاهده شده در تست‌های جفت فقط به خاطر 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			
	90%	95%	98%	99%
υ=2	2.92	4.30	6.96	9.92
υ=5	2.02	2.57	3.36	4.03
υ=10	1.81	2.23	2.76	3.17
υ=20	1.72	2.09	2.53	2.84
υ=30	1.70	2.04	2.46	2.75
υ=120	1.66	1.98	2.36	2.62
υ=∞	1.64	1.96	2.33	2.58

جدول ۵،۶ مقادیر $\boldsymbol{t_{N,v}}$ برای بازه‌های اطمینان دو طرفه. با $v\rightarrow\infty$، $\boldsymbol{t_{N,v}}$ به $\boldsymbol{z_N}$ میل می کند.

[۸] number of degrees of freedom
[۹] paired tests

۵,۶,۱ تست‌های جفتی t[۱۰]

در بالا، فرایندی برای مقایسه‌ی دو متد یادگیری با مجموعه‌ی ثابتی از داده‌ها ارائه شد. در این بخش توجیه آماری برای ایـن فراینـد و بازه‌هـای اطمینان روابط ۵,۱۷ و ۵,۱۸ را مورد بحث قرار می‌دهیم. در اولین بار خواندن این کتاب می‌توانید بـدون از دسـت دادن پیوسـتگی مطالـب ایـن قسمت را نخوانید.

بهترین راه برای درک توجیه بازه‌های تخمینی اطمینان که در رابطه‌ی ۵,۱۷ 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 زننده‌ی 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، بسـیار کلـی اسـت. بـرای مثـال، ایـن حالـت کلـی مسـئله‌ی تخمـین $error_{\mathcal{D}}(h)$ با استفاده از $error_S(h)$ را نیز شامل می‌شود. (در این مسئله، Y_iها ۱ یا ۰ بودند (بنا بـه اینکـه نمونـه درسـت دسـته‌بندی می‌شود یا خیر) و $error_{\mathcal{D}}(h)$ همان میانگین بود که می‌خواستیم تخمین بزنیم). تست **t**، که در رابطه‌ی ۵,۱۷ و ۵,۱۸ نیز آورده شده اسـت، 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 ۵,۵ را برای مقایسه‌ی کارایی دو الگوریتم یادگیری در نظر بگیرید. فـرض کنیـد کـه بـه جـای داشـتن مجموعه‌ی ثابت D_0، می‌توانیم نمونه‌های آموزشی جدید را بر اساس توزیع احتمال $\mathcal{D}$ دریافت کنیم. در کل، متد ایده آل فراینـد جـدول ۵,۵ در هر بار تکرار حلقه از یک دسته‌ی آموزشی جدید S_{i} و دسته‌ی تست جدید $\mathrm{T_i}$ را که با توزیع $\mathcal{D}$ (همان توزیع D_0) ایجاد شده استفاده مـی‌کنـد. این متد ایده آل کاملاً با فرم مسئله‌ی تخمینی بالا تطبیق دارد. در کل، معیار δ_{i} در این فرایند متناسب با متغیرهـای تصـادفی بـه طـور یکسـان توزیع شده Y_{i} ها هستند. میانگین μ این توزیع‌ها مقدار امید اختلاف بین خطاها بین دو متد یادگیری را نشان می‌دهد (رابطـه‌ی ۵,۱۴). میـانگین نمونه‌ای $\bar{\mathrm{Y}}$ کمیتی است $\bar{\delta}$ که توسط حالت ایده آل این متد اندازه‌گیری می‌شود. علاقه‌ی ما به جواب این سؤال است کـه "میـانگین $\bar{\delta}$ تـا چـه 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 ۳۰ نمونـه داشـته باشـد. بـه همین دلیل، هر کدام از δ_i ها (طبق قضیه‌ی حد مرکزی) توزیعی تقریباً نرمال خواهند داشت. بنابراین، با حالت خاصی روبرو هسـتیم کـه 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، توزیـع 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 ۵,۱۱ که برای توزیع احتمال‌های با توزیع نرمال صادق بود). متأسفانه، این رابطه نیاز بـه انحراف معیار دارد، که در حال حاضر برای ما متغیری مجهول است.

[۱۰] Paired t tests

تست t دقیقاً برای چنین شرایطی به وجود آمده است، شرایطی که در آن هدف تخمین میانگین نمونه‌ای مجموعه‌ای از متغیرهـای تصـادفی بـا توزیع‌های نرمال به طور یکسان توزیع شده است. در چنین شرایطی، می‌توان از بازه‌های اطمینان رابطه‌ی ۵٫۱۷ و ۵٫۱۸ که با نمادگذاری جدیـد به شکل زیر نمایش داده می‌شوند استفاده کرد:

$$\mu = \bar{Y} \pm \mathrm{t}_{\mathrm{N,k-1}} \mathrm{s}_{\bar{Y}}$$

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

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

و $\mathrm{t}_{\mathrm{N,k-1}}$ ثابتی مشابه ثابت قبلی z_N است. در واقع ثابت $\mathrm{t}_{\mathrm{N,k-1}}$ ثابتی است که ویژگی‌های ناحیه‌ای از توزیع احتمال موسوم به توزیـع t را مشابه ثابت z_N که ناحیه‌ای از توزیع احتمال نرمال را مشخص می‌کرد مشخص می‌کند. توزیع t مشابه توزیع نرمال توزیعی زنگی شکل است با این تفاوت که پهنای بیشتری دارد تا واریانس $\mathrm{s}_{\bar{Y}}$ 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 $\mathrm{t}_{\mathrm{N,k-1}}$ نیز به z_N) میل خواهد کرد. این منطقی است زیرا که انتظار داریم که $\mathrm{s}_{\bar{Y}}$، 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.

۵٫٦٫۲ نکات کاربردی

توجه دارید که بحث بالا استفاده از تخمین بازه‌ی اطمینان رابطه‌ی ۵٫۱۷ را در حالتی که علاقه ما به میانگین نمونه‌ای $\bar{Y}$ برای تخمین میـانگین 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 اضـافه شـده. در عمـل، بـا داشـتن مجموعـه‌ی محـدود D_0 از نمونه‌های آموزشی و متد عملی جدول ۵٫۵، این توجیه کاملاً برقرار نیست. در کل، مسئله اینجاست که تنها راه ایجاد δ_i های جدید بـاز ترکیـب 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 ۵٫۵ متدی به نام k-fold را که در آن مجموعه‌ی D_0 را به k زیرمجموعه‌ی هم‌اندازه تقسیم مـی‌کنـد. در ایـن روش، هـر نمونـه‌ی D_0 دقیقـاً در یـک مجموعه‌ی تست استفاده و k-1 بار به عنوان نمونه‌ی آموزشی مورد استفاده قرار می‌گیرد. راه‌حل دیگر متداول انتخاب تصادفی حداقل ۳۰ نمونه از D_0 به عنوان مجموعه‌ی تست و استفاده از بقیه‌ی نمونه‌ها برای آموزش است، این متد را می‌توان بـه تعـداد دلخـواه تکـرار کـرد. ایـن متـد تصادفی این مزیت را دارد که می‌توان برای کوچک کردن بازه‌های اطمینان به اندازه‌ی دلخواه، آن را بی‌نهایت بار تکرار کرد. در مقابل، متـد 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 بایـد بیشتر ۳۰ 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، دسته‌های تست ایجاد شده در روش k-fold از یکدیگر مستقل خواهند بود زیرا که هر نمونه تنها در یک دسته‌ی تست حضور دارد.

به طور خلاصه، هیچ فرایندی در مقایسه‌ی متدهای یادگیری بر اساس داده‌های محدود تمامی ویژگی‌هایی که ما می‌خواهیم را ندارد. پس باید در نظر داشت که مدل‌های آماری در تست الگوریتم‌های یادگیری زمانی که تعداد داده‌های موجود محدود است به ندرت تمامی ویژگی‌های مورد نظر را خواهند داشت. با این وجود، این مدل‌ها بازه‌های اطمینان را که می‌توانند کمک بزرگی در تفسیر آزمایش‌های مقایسه‌ی متدهای یادگیری است را ارائه می‌کنند.

۵,۷ خلاصه و منابع برای مطالعه‌ی بیشتر

نکات اصلی این فصل شامل موارد زیر می‌شود:

- نظریه‌ی آمار مبنایی برای تخمین از خطای واقعی ($error_D(h)$) فرضیه‌ی h بنا بر مشاهداتش از خطای مشاهده شده ($error_S(h)$) بر روی نمونه‌ی S ارائه می‌کند. برای مثال، اگر h یک فرضیه‌ی گسسته مقدار باشد و تعداد داده‌های نمونه‌ی S بیش از ۳۰ باشد که به طور مستقل از یکدیگر انتخاب شده‌اند، آنگاه بازه‌ی اطمینان N% برای خطای $error_D(h)$ تقریباً بازه‌ی زیر خواهد بود:

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

در این رابطه مقدار z_N از جدول ۵,۱ تعیین می‌شود.

- در کل، مشکل تخمین بازه‌ی اطمینان با تعیین پارامتری که باید تخمین زده شود ($error_D(h)$) و یک تخمین زننده ($error_S(h)$) برای این کمیت انجام می‌گیرد. چون تخمین زننده یک متغیر تصادفی است ($error_S(h)$ وابسته به مجموعه نمونه‌ی تصادفی S است)، آن را می‌توان با تابع توزیع احتمال حاکم نشان داد. بازه‌های اطمینان را می‌توان با پیدا کردن بازه‌ای 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 زننده‌ی پارامتر p باشد، بایاس تخمین Y خطای بین p و مقدار امید Y خواهد بود. برای مثال اگر S داده‌های آموزشی برای ساخت فرضیه‌ی 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 زننده‌ی بدون بایاس نیز مقدار مشاهده شده تخمین زننده در آزمایش‌های متفاوت با هم متفاوت است. واریانس σ^2 ی توزیع حاکم بر خواص تخمین زننده تعیین می‌کند که این مقدار از مقدار واقعی چقدر می‌تواند متفاوت باشد. این واریانس با افزایش تعداد نمونه‌های داده کاهش می‌یابد.
- مقایسه‌ی کارایی دو الگوریتم یادگیری نیز یک مسئله‌ی تخمین است، که آن زمانی که زمان و داده‌های آموزشی نامحدودند، بسیار ساده است اما هنگامی که منابع محدود می‌شوند این مسئله کمی سخت‌تر می‌گردد. یکی از راه‌های 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 بازه‌های 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 ۵,۲ 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

۵,۱ 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 ۵,۳,۴ مقایسه کرد؟

۵,۲ فرضیه‌ی یاد گرفته شده‌ی h برای مفهومی منطقی را در نظر بگیرید. هنگامی که h بر روی مجموعه‌ای از ۱۰۰ نمونه بررسی می‌شود ۸۳تای آن‌ها را درست دسته‌بندی می‌کند. انحراف معیار و بازه‌ی ۹۵٪ اطمینان را برای خطای $error_D(h)$ بیابید.

۵,۳ فرض کنید که فرضیه‌ی h بر روی نمونه‌ای مستقل با n=65 دارای خطای r=10 است. بازه‌ی دوطرفه‌ی ۹۰٪ اطمینان برای خطای واقعی چقدر است؟ بازه‌ی ۹۵٪ اطمینان یک طرفه چقدر است (یعنی با احتمال ۹۵٪ داریم $error_S(h) \leq U$)؟ بازه‌ی اطمینان ۹۰٪ درصد یک طرفه چقدر است؟

۵,۴ 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 ۵,۵ را تغییر دهید)

۵,۵ توضیح دهید که چرا تخمین بازه‌ی اطمینان رابطه‌ی ۵,۱۷ به تخمین کمیت رابطه‌ی ۵,۱۶ اعمال می‌شود و چرا نمی‌توان آن را به ۵,۱۴ 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 نمونه‌برداری
sample error	خطای نمونه‌ای
true error	خطای واقعی
well suited	خوش‌تعریف
number of degrees of freedom	درجه‌ی آزادی
central limit	قضیه‌ی حد مرکزی
two sided bound	مرز دو طرفه
one side bound	مرز یک طرفه