کنترل‌کردن سوییدگی در استفاده‌کردن از اطّلاعات

How much does your data exploration overfit?
Controlling bias via information usage
Daniel Russo and James Zou

گزارش پروژه‌ی درس تئوری اطّلاعات

بهراد منیری   محمّدرضا رحمانی

فهرست مطالب

۱ مقدّمه

در این پروژه به بررسی مقاله‌ی [۱] می‌پردازیم. در تحلیل داده‌های با بُعد بالا، عموماً آنالیز که باید بر روی داده انجام شود و متغیرهایی که باید تخمین زده شوند، قبل از مشاهده و بررسی ابتدایی داده‌ها برای تحلیل‌گر واضح نیستند. به همین سبب، در اکثر مواقع، تحلیل به صورت تطبیقی انجام می پذیرد، به این نحو که با انجام آزمایش‌هایی بر روی داده و به صورت تدریجی، آنالیزهای جالبی که باید بر روی داده انجام شوند کشف شده و سپس به کمک همین داده، این آنالیزها انجام می‌پذیرند. این تحلیل تطبیقی و مشاهده‌ی داده، قبل از انجام آنالیز نهایی، ممکن است باعث ایجاد سوییدگی یا بایاس در نتیجه‌ی تحلیل نهایی شود. در آمار و استنتاج آماری، معمولاً فرض بر این است که قبل از انجام هر تحلیل، به داده هیچ «نگاهی» نشده است و انتخاب تحلیل‌های انجام شده مستقل از داده هستند. این فرض اساسی در آمار کلاسیک، تفاوتی چشمگیر با آنچه در عمل تحلیل‌گران داده انجام می‌دهند دارد. به عنوان مثال فرض کنید تحلیل گر داده قصد دارد داده‌های خود را به کمک الگوریتمی ساده مثل k-means به تعدادی خوشه تقسیم کند. معمولاً تعداد خوشه‌ها توسط تحلیل‌گر و بعد از رسم نمودارهای مختلفی از داده تعیین می‌شود. آیا این استفاده از داده برای تعیین آنالیز انجام شده، می تواند باعث خطای بزرگی در نتایج تحلیل شود؟ آیا راهی برای کمّی‌سازی این بایاس و کنترل آن وجود دارد؟ در این مقاله به چنین سوالاتی پاسخ داده می‌شود.

در این مقاله، هدف ارائه‌ی چهارچوبی نظریه اطّلاعاتی برای کران‌زدن سوییدگی ناشی از نشت اطّلاعات داده در حین فرآیند انتخاب یک آنالیز است. در این چهارچوب، بر خلاف استنتاج کلاسیک آماری، فرض بر این است که از خود داده در فرآیند تعیین تحلیل استفاده شده است. این مقاله به کمک چهارچوب ارائه‌شده، سوییدگی تعدادی از فرآیندهای معروف تحلیل داده‌ای که از خود داده برای انتخاب تحلیل مورد نظر استفاده می‌کنند را مطالعه می‌کند. در آخر نیز از ایده‌ی اضافه کردن نویز در حین تحلیل داده برای کاهش سوییدگی استفاده کرده و به بررسی دسته‌ای از فرآیندهای تطبیقی تحلیل داده می‌پردازد.

۱.۱ نمونه‌هایی از تحلیل تطبیقی داده

در این بخش، دو مثال از تحلیل تطبیقی داده ارائه می‌شود.

مثال ۱.۱. بهراد دیتاستی مربوط به تغییرات وزن و علل ژنتیکی آن در اختیار دارد. این دیتاست شامل داده‌ی ۱۰۰۰ نفر است و برای هر شخص، وزن در سه زمان مشخص اندازه گیری شده است. در این دیتاست، برای هر فرد نیز *expression value* ۲۰۰۰ ژن نیز ثبت گشته است. سه تغییر وزن در این دیتاست می‌تواند مورد بررسی قرار بگیرد. تغییر وزن از زمان ۱ تا ۲، زمان ۲ تا ۳ یا از زمان ۱ تا ۳. بهراد، قبل از شروع هر آزمایشی، تصمیم گرفته است که تنها تغییرات از زمان ۱ تا ۳ را مورد بررسی قرار دهد. بهراد مقدار همبستگی هر ژن با تغییر وزن را محاسبه کرده و ژن با بیشترین همبستگی، به همراه *R-Squared* رابطه‌ی خطی تغییر وزن و آن میزان *expression* ژن را گزارش می‌کند. مشاهده می‌شود که اگر بار دیگر این آزمایش به صورت مستقل انجام شود، همبستگی ویژگی انتخاب شده توسط بهراد با تغییر وزن در داده‌های جدید، کمتر از همبستگی به دست آمده توسط بهراد در این آزمایش خواهد بود. این پدیده حتی در صورتی که بهراد ابتدا به کمک یک آزمون فرضیه، تعدادی از ژن‌ها را که همبستگی آن‌ها با وزن معنادار است انتخاب کرده و سپس ماکزیمم همبستگی را در بین آنها انتخاب کند نیز مشاهده می‌شود. این پدیده، *Winner's Curse Selection Bias* نام دارد. این سوییدگی، به این دلیل ایجاد شده است که بهراد از همین دیتاست برای انتخاب ژن مدنظر استفاده کرده است.

مثال ۲.۱. محمّدرضا همان داده‌ی بهراد را در اختیار دارد و چند آزمایش ساده با آن انجام می‌دهد. در ابتدا، برای هر زمان، میانگین *expression* ژن‌ها در افراد مختلف محاسبه می‌کند. مشاهده می‌کند که این میانگین در زمان ۱ و ۲ تفاوت چندانی ندارد ولی در زمان ۳ به مراتب بزرگ‌تر شده است، بنابر این توجه خود را به زمان ۲ و ۳ جلب می‌کند. همچنین او مشاهده می‌کند که نیمی از ژن‌ها همواره *expression* کمی دارند و لذا محمّدرضا آن‌ها را به راحتی حذف می‌کند و تنها ۱۰۰۰ ژن را نگه می‌دارد. در آخر نیز از این ۱۰۰۰ ژن، ژنی را معرفی می‌کند که بیشترین همستگی با تغییرات وزن در بازه‌ی زمانی ۲ تا ۳ را دارد. تحلیل محمّدرضا به شدت تطبیقی است و نتایج کلاسیک آمار به راحتی قابل استفاده برای آن نیست. حدس می‌زنیم که اگر آزمایش مجدد تکرار شود، ژن انتخاب شده همبستگی کمتری با چیزی که محمدرضا به دست آورده داشته باشد. چگونه می‌توانیم این حدسیات را کمّی کنیم؟

این دو مثال، نمونه‌هایی از تحلیل‌های تطبیقی و سوییدگی ناشی از آن‌ها را در پردازش داده معرفی می‌کند. در این مثال‌ها می‌توان به دو شهود از این سوییدگی رسید. اول این‌که به نظر می‌رسد که این سوییدگی به توزیع خود داده ربط داشته باشد. به عنوان مثال، در تحلیل بهراد اگر یکی از ژن‌ها همبستگی خیلی بیشتر به نسبت سایر ژن‌ها داشته باشد، شهوداً با احتمال بالایی در هر تکرار مستقل آزمایش، همین ژن انتخاب می‌شود و مقدار سوییدگی در این حالت کم است. دوم این‌که مقدار سوییدگی ناشی از مرحله از

تحلیل می‌تواند با سوییدگی ناشی از مراحل دیگر متفاوت باشد. به عنوان مثال در تحلیل محمّدرضا، شهوداً مرحله‌ی انتخاب ژن با بیشترین expression در مرحله‌ی آخر، به نسبت سایر مراحل مولّد سوییدگی بیشتری است.

۲.۱ مدل ریاضی تحلیل تطبیقی داده

در مدل این مقاله از تحلیل داده، دیتاست $D \in \mathcal{D}$ با توزیع $\mathcal{P}$ از فضای تمام دیتاست‌های ممکن، $\mathcal{D}$ انتخاب شده است. تحلیل‌گر داده، تعداد زیاد m «تحلیل» مختلف را قبل از مشاهده‌ی دیتاست در نظر دارد، اما قصد دارد نتایج تنها یکی از آن‌ها راگزارش کند و این آنالیز را بعد از مشاهده‌ی D و یا آماره‌هایی از D انتخاب می‌کند. به صورت دقیق‌تر، تحلیل‌گر تعداد m تابع $\phi_1, \phi_2, \ldots, \phi_m$ را در نظر گرفته که $\phi_i : \mathcal{D} \to \mathbb{R}$ است. بعد از مشاهده‌ی داده‌ها، تحلیل‌گر تابع $\phi_{T(D)}(D)$ را انتخاب کرده و آن را گزارش می‌کند. از آنجایی که خود T تابعی از D است، ممکن است سوییدگی بزرگی در انتخاب $\phi_{T(D)}(D)$ وجود داشته باشد. به عنوان مثال اگر تمام ϕ_i ها دارای میانگین صفر باشد، ممکن است $\mathbb{E}[\phi_{T(D)}]$ به طرز معناداری مثبت باشد. در ادامه به بیان دو مثال (۱.۱) و (۲.۱) در قالب فرمول‌بندی مطرح شده خواهیم پرداخت.

مثال ۳.۱. در مثال (۱.۱)، دیتاست $D \in \mathbb{R}^{1000 \times 2003}$ است و شامل ۲۰۰۰ ویژگی و علاوه‌ی وزن در سه زمان مختلف است. این ویژگی‌ها نیز برای هزار نفر ثبت شده‌اند. توابع ϕ_i نیز همبستگی بین ژن‌ها و تغییر وزن از زمان ۱ تا ۳ هستند. بهراد نیز در نهایت $T = \text{argmax}_i \phi_i$ را انتخاب می‌کند.

مثال ۴.۱. در مثال (۲.۱)، محمّدرضا همان دیتاست بهراد است. توابع ϕ_i همبستگی هر ژن با ۳ تغییر وزن هستند. انتخاب T نیز به شیوه‌ی پیچیده‌ای انجام شده است.

میانگین واقعی تحلیل i ام را $\mu_i = \mathbb{E}[\phi_i]$ بنامید. برای یک دیتاست مشخص D، اگر $T(D) = i$ باشد، خروجی برابر ϕ_T است. خروجی واقعی نیز μ_T است. مقدار $\phi_T - \mu_T$ خطای ناشی از تحلیل تطبیقی داده است. سوییدگی تحلیل را برابر امیدریاضی این کمیت تعریف می‌کنیم، یعنی $\mathbb{E}[\phi_T - \mu_T]$ که در آن، امید ریاضی بر روی عدم قطعیت دیتاست D و روش انتخاب $T(D)$ است.

در این مقاله، کرانی برای $\mathbb{E}[\phi_T - \mu_T]$ بر مبنای «استفاده‌ی بد از اطّلاعات» در انتخاب T معرفی می‌شود و نشان داده‌می‌شود که اگر در انتخاب T، از اطّلاعاتی از داده استفاده شود که در ϕ_i ها وجود ندارد، یعنی تنها از «اطّلاعات خوب» استفاده شده باشد، این تحلیل تطبیقی منجر به سوییدگی نخواهد شد و سوییدگی تنها در حالتی رخ می‌دهد که در انتخاب T از اطّلاعات کدشده در ϕ_i ها استفاده شود.

۳.۱ کارهای پیشین

در این بخش به بررسی اجمالی ادبیات پیشین این حوزه می‌پردازیم. خط مهمی از پژوهش در تئوری یادگیری ماشین، بررسی مفهوم پایداری الگوریتمی و استفاده از آن برای جلوگیری از over-fitting است [۲، ۳، ۴]. چهارچوب یادگیری PAC-Bayes نیز مفهومی مرتبط است که به ارائه‌ی کران‌های تعمیم الگوریتم‌های یادگیری ماشین بر حسب فاصله‌ی KL می‌پردازد [۵].

تفاوت جدی کار حاضر با کارهای پیشین این است که در روش‌های گذشته، مبتنی بر پایداری یا تئوری PAC، کران‌ها برای بدترین حالت ارائه شده و برای هر توزیع ورودی دلخواهی برقرار هستند و به طور مثال، پایداری یک الگوریتم ربطی به توزیع داده‌ای که با آن داده می‌شود ندارد. روش نظریه‌ی اطّلاعاتی حاضر، از آن‌جا که بدترین حالت را در نظر نمی‌گیرد، قادر است کران‌های بهتری ارائه کند.

۲ کنترل‌کردن سوییدگی از راه محدودکردن استفاده از اطّلاعات

۱.۲ معرّفی یک کران بالا برای سوییدگی به وسیله‌ی اطّلاعات استفاده شده

در این بخش، به معرّفی یک کران بالا برای سوییدگی با استفاده از یک کمیّت مورد استفاده در نظریه‌ی اطّلاعات می‌پردازیم. فرض کنید مجموعه‌داده‌ی D داده شده است و متغیّرهای $\{\phi_1(D), \phi_2(D), \cdots, \phi_m(D)\}$ محاسبه شده‌اند. ما یک اندیس $T(D)$ را انتخاب می‌کنیم و $\phi_T(D)$ را گزارش می‌کنیم. سؤال آنست که مقدار سوییدگی این انتخاب $(\phi_T(D) - \mathbb{E}[\phi_T(D)])$ چقدر است؟

برای این کار، فرض کنید که Ω، فضای نمونه‌ای مجموعه‌داده‌ی D باشد و $\boldsymbol{\phi} = (\phi_1, \phi_2, \cdots, \phi_m) : \Omega \to \mathbb{R}^m$ و $T : \Omega \to \{1, 2, \cdots, m\}$ دو متغیّر تصادفی باشند که روی فضای نمونه‌ای مشترک Ω تعریف شده‌اند. همچنین فرض کنید $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_m) \triangleq \mathbb{E}[\boldsymbol{\phi}]$ امید ریاضی بردار تصادفی $\boldsymbol{\phi}$ باشد. در صورتی که برای هر $i \in \{1, 2, \cdots, m\}$، $\phi_i - \mu_i$ یک متغیّر تصادفی زیر–گاوسی باشد، می‌توان کران بالایی برای سوییدگی با استفاده از اطّلاعات متقابل میان T و $\boldsymbol{\phi}$ یافت.

قضیه ۱.۲. متغیّر تصادفی برداری $\boldsymbol{\phi} = (\phi_1, \cdots, \phi_m)$ را در نظر بگیرید، اگر تعریف کنیم

$$\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_m) = \mathbb{E}[\boldsymbol{\phi}]$$

و اگر برای هر $i \in \{1, \cdots, m\}$، $\phi_i - \mu_i$ یک متغیّر تصادفی زیر–گاوسی با پارامتر σ باشد، آن‌گاه:

$$|\mathbb{E}[\phi_T - \mu_T]| \leq \sigma\sqrt{2I(T;\boldsymbol{\phi})} \tag{۱}$$

کمیّت $I(T;\boldsymbol{\phi})$ را «اطّلاعات استفاده‌شده» می‌نامیم و بیان‌گر میزان وابستگی اندیس انتخابی T به مقادیر $(\phi_1, \cdots, \phi_m)$ است. از نظر شهودی، می‌توان ϕ_iها را تخمین‌هایی از یک کمیت دانست و در این صورت، $I(T;\boldsymbol{\phi})$ وابستگی T به نویز موجود در این تخمین‌ها را نشان می‌دهد. همچنین می‌توان $I(T;\boldsymbol{\phi})$ را «استفاده‌ی بد از اطّلاعات نامید» که بیان‌گر آنست که چه میزان از نویز موجود در داده‌ها $(D \sim \mathcal{P})$ در انتخاب تخمین گزارش‌شده اثر می‌گذارد. به تعبیر دیگر، اگر ما به طمع رسیدن به بیشترین اطّلاعات، T را به شدّت به ϕ_iها وابسته کنیم، هم‌زمان تأثیر اطّلاعات بد در T انتخابی را هم افزایش داده‌ایم و این می‌تواند باعث ایجاد overfitting شود.

وقتی که T به طور کامل توسّط ϕ_iها تعیین شود، $I(T;\boldsymbol{\phi}) = H(T)$ می‌شود که به معنای آنست که در تحقّق‌های مختلف داده‌ها، T چگونه تغییر می‌کند.

در صورت قضیه‌ی (۱.۲)، فرض شده‌است که همه‌ی $\phi_i - \mu_i$ها، زیر–گاوسی با پارامتر مشترک $\sigma_i = \sigma$ هستند. در حالتی که σ_iها برابر نباشند، یک تعمیم از قضیه‌ی (۱.۲) می‌تواند به صورت $E[\phi_T - \mu_T]| \leq \max_i\{\sigma_i\}\sqrt{2I(T;\boldsymbol{\phi})}$ بیان شود، ولی کران بهتری هم وجود دارد که در قضیه‌ی (۲.۲) بیان شده‌است.

قضیه ۲.۲. متغیّر تصادفی برداری $\boldsymbol{\phi} = (\phi_1, \cdots, \phi_m)$ را در نظر بگیرید، اگر تعریف کنیم $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{\phi}]$ و اگر برای هر $i \in \{1, \cdots, m\}$، $\phi_i - \mu_i$ یک متغیّر تصادفی زیر–گاوسی با پارامتر σ_i باشد، آن‌گاه:

$$|\mathbb{E}[\phi_T - \mu_T]| \leq \sqrt{\mathbb{E}[\sigma_T^2]}\sqrt{2I(T;\boldsymbol{\phi})} \tag{۲}$$

همچنین اگر فرض زیر–گاوسی بودن $\phi_i - \mu_i$ها را کمی ضعیف کنیم و به زیر–نمایی بودن تقلیل دهیم، می‌توان قضیه‌ی مشابهی را بیان کرد.

قضیه ۳.۲. متغیّر تصادفی برداری $\boldsymbol{\phi} = (\phi_1, \cdots, \phi_m)$ را در نظر بگیرید، اگر تعریف کنیم $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{\phi}]$ و اگر برای هر $i \in \{1, \cdots, m\}$، $\phi_i - \mu_i$ یک متغیّر تصادفی زیر–نمایی با پارامترهای (σ, b) باشد، آن‌گاه:

$$\mathbb{E}[\phi_T - \mu_T] \leq bI(T;\boldsymbol{\phi}) + \frac{\sigma^2}{2b} \tag{۳}$$

علاوه بر این، اگر $b < 1$ باشد، داریم:

$$\mathbb{E}[\phi_T - \mu_T] \leq \sqrt{b}I(T;\boldsymbol{\phi}) + \frac{\sigma^2}{2\sqrt{b}} \tag{۴}$$

برای فهم بهتر قضیه‌ی (۱.۲)، چند مثال را در نظر می‌گیریم.

مثال ۱.۲. فرض کنید $\phi_i = \frac{1}{n}\sum_{i=1}^{n} f_i(X_i)$ که داده‌های X_i به صورت $i.i.d.$ از توزیع $p_X(x)$ برداشته شده‌اند. در این صورت، اگر $f_i(X_j) - \mathbb{E}[f_i(X_j)]$ یک متغیّر تصادفی زیر–گاوسی با پارامتر σ باشد، $\phi_i - \mu_i$ یک متغیّر تصادفی زیر–گاوسی با پارامتر $\frac{\sigma}{\sqrt{n}}$ است و در نتیجه:

$$|\mathbb{E}[\phi_T - \mu_T]| \leq \sigma\sqrt{\frac{2I(T;\boldsymbol{\phi})}{n}} \tag{۵}$$

مثال ۲.۲. فرض کنید T مستقل از $\boldsymbol{\phi}$ انتخاب شود، این حالت وقتی رخ می‌دهد که انتخاب T زودتر از مشاهده‌ی داده‌ها رخ داده و نمی‌تواند با مشاهده‌ی داده‌ها و ϕ_iها تغییر کند. همچنین اگر داده‌ها را به دو بخش مستقل تقسیم کنیم و از یک بخش برای تعیین T استفاده کنیم و سپس ϕ_T را با استفاده از بخش دوم داده‌ها محاسبه و گزارش کنیم، در این حالت هم T از $\boldsymbol{\phi}$ مستقل است. در این حالت $I(T;\boldsymbol{\phi})=0$ و در نتیجه $\mathbb{E}[\phi_T]=\mathbb{E}[\mu_T]$.

مثال ۳.۲. فرض کنید که هرکدام از ϕ_iها، یک متغیّر تصادفی گاوسی با توزیع $\mathcal{N}(0,\sigma^2)$ باشند. اگر فرض کنیم $T=\arg\max_{1\leq i\leq m}\phi_i$، در این صورت داریم $I(T;\boldsymbol{\phi})=H(T)=\log(m)$، زیرا ϕ_iها تقارن دارند و با احتمال برابر، هرکدام می‌توانند از بقیه بیشتر باشند. در نتیجه داریم:

$$\mathbb{E}[\phi_T-\mu_T]=\mathbb{E}[\phi_T]\leq\sigma\sqrt{2\log(m)} \tag{۶}$$

می‌توان نشان داد که با افزایش m، این نامساوی به تساوی تبدیل می‌شود.

یک حالت کلّی‌تر را نیز می توان در نظر گرفت. فرض کنیم ابتدا ϕ_iها از بزرگ‌ترین تا کوچک‌ترین مرتّب می‌شوند، و سپس یکی از $m_0<m$ اندیس مربوط به ϕ_iهای بزرگ‌تر، با توزیع یکنواخت انتخاب می‌شود. در این حالت، داریم $I(T;\boldsymbol{\phi})=H(T)-H(T|\boldsymbol{\phi})$ و به دلیل تقارن ϕ_iها، $H(T)=\log(m)$ و به دلیل توزیع یکنواخت T روی m_0 اندیس مربوط به ϕ_iهای بزرگ‌تر، $H(T|\boldsymbol{\phi})=\log(m_0)$. در نتیجه:

$$\mathbb{E}[\phi_T-\mu_T]=\mathbb{E}[\phi_T]\leq\sigma\sqrt{2\log(\frac{m}{m_0})} \tag{۷}$$

۲.۲ اطّلاعات استفاده شده، به عنوان کرانی برای بقیه‌ی معیارهای سوییدگی

در بخش قبل، مشاهده کردیم که اطّلاعات استفاده شده می‌تواند به عنوان کران بالایی برای $|\mathbb{E}[\phi_T-\mu_T]|$ به کار رود. گاهی وقت‌ها ما به معیارهای دیگری برای سوییدگی، نیاز داریم، مانند $\mathbb{E}[|\phi_T-\mu_T|]$ و $\mathbb{E}\left[(\phi_T-\mu_T)^2\right]$. در این بخش می‌خواهیم به کمک $\sqrt{I(T;\boldsymbol{\phi})}$ و $I(T;\boldsymbol{\phi})$، کرانی برای بقیه‌ی معیارهای سوییدگی هم بیابیم.

قضیه ۴.۲. متغیّر تصادفی برداری $\boldsymbol{\phi}=(\phi_1,\cdots,\phi_m)$ را در نظر بگیرید، اگر تعریف کنیم $\boldsymbol{\mu}=\mathbb{E}[\boldsymbol{\phi}]$ و اگر برای هر $i\in\{1,\cdots,m\}$، $\phi_i-\mu_i$ یک متغیّر تصادفی زیر–گاوسی با پارامتر σ باشد، آن‌گاه:

$$\mathbb{E}[|\phi_T-\mu_T|]\leq\sigma+c\sigma\sqrt{2I(T;\boldsymbol{\phi})} \tag{۸}$$

که $c<36$ یک ثابت است.

قضیه ۵.۲. متغیّر تصادفی برداری $\boldsymbol{\phi}=(\phi_1,\cdots,\phi_m)$ را در نظر بگیرید، اگر تعریف کنیم $\boldsymbol{\mu}=\mathbb{E}[\boldsymbol{\phi}]$ و اگر برای هر $i\in\{1,\cdots,m\}$، $\phi_i-\mu_i$ یک متغیّر تصادفی زیر–گاوسی با پارامتر σ باشد، آن‌گاه:

$$\mathbb{E}\left[(\phi_T-\mu_T)^2\right]\leq 1.25\sigma^2+c_2\sigma^2 I(T;\boldsymbol{\phi}) \tag{۹}$$

که $c\leq 10$ یک ثابت است.

۳.۲ اطّلاعات استفاده شده به عنوان یک کران پایین برای سوییدگی

در این بخش می‌خواهیم از اطّلاعات متقابل به عنوان کران پایینی برای سوییدگی استفاده کنیم.

فرض کنید $\boldsymbol{\phi}\sim\mathcal{N}(\boldsymbol{\mu},I)$ و $T=\arg\max_i\phi_i$، از آنجا که T یک تابع یقینی از $\boldsymbol{\phi}$ است، $I(T;\boldsymbol{\phi})=H(T)$. در نتیجه از قضیه‌ی (۵.۲) می‌توان نوشت:

$$\mathbb{E}\left[(\phi_T-\mu_T)^2\right]\leq\sigma^2(1.25+10I(T;\boldsymbol{\phi}))=\sigma^2(1.25+10H(T)) \tag{۱۰}$$

در این قسمت نشان می‌دهیم که $H(T)$ می‌تواند به عنوان یک کران پایین از $\mathbb{E}\left[(\phi_T-\mu_T)^2\right]$ هم به کار رود.

قضیه ۶.۲. فرض کنید $T = \arg\max_i \phi_i$ که در آن، $\boldsymbol{\phi} \sim \mathcal{N}(\boldsymbol{\mu}, I)$. در این صورت ثابت‌های $c_1 = \frac{1}{8}$، $c_2 < 2.5$، $c_3 = 10$ و $c_4 = 1.5$ وجود دارند، به قسمی که:

$$c_1 H(T) - c_2 \leq \mathbb{E}\left[(\phi_T - \mu_T)^2\right] \leq c_3 H(T) + c_4 \tag{۱۱}$$

برای به دست آوردن یک دید شهودی از این رابطه، یک حالت ساده را در نظر می‌گیریم: فرض کنید $m = 2$، $\phi_1 = x$ که x یک مقدار ثابت است و $\phi_2 \sim \mathcal{N}(0,1)$. همچنین فرض کنید $T = \arg\max_i \phi_i$. اگر $x \gg 0$ باشد، داریم:

$$\log(\frac{1}{\mathbb{P}[T=2]}) = \log(\frac{1}{\mathbb{P}[\phi_2 \geq x]}) \approx \frac{x^2}{2}$$

در نتیجه در حالتی که $x \to \infty$ می‌توان نوشت:

$$H(T_x) \sim \mathbb{P}[T_x = 2] \log(\frac{1}{\mathbb{P}[T_x = 2]}) \sim \mathbb{P}[T_x = 2] x^2 \sim \mathbb{E}\left[(\phi_{T_x} - \mu_{T_x})^2\right]$$

که T_x به معنای اندیس انتخاب شده است، وقتی که $\phi_1 = x$ باشد و $f(x) \sim g(x)$ به معنای آنست که اگر $x \to \infty$، آنگاه $\frac{f(x)}{g(x)} \to 1$

۳ چه زمانی سوییدگی بزرگ است؟

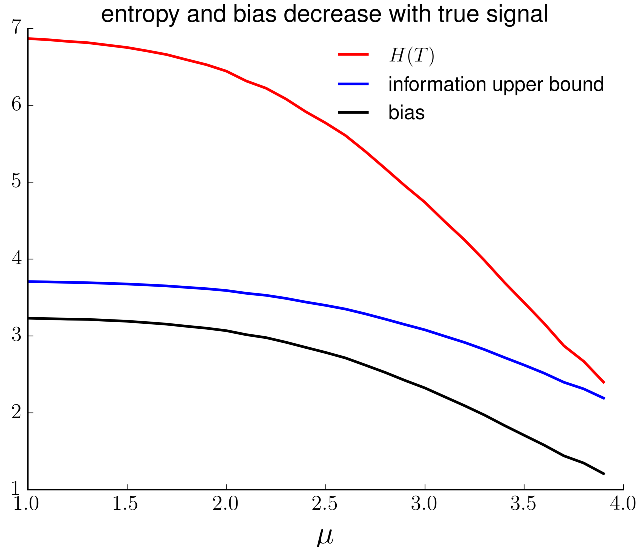
در این بخش، به بررسی چند روش ساده، ولی معمول و پر استفاده در انتخاب ویژگی و تخمین پارامتر می‌پردازیم. در بسیاری از کاربردها، انتخاب تحلیل و تخمین آن بر روی همان دیتاست اصلی انجام می‌شوند. روش بهره‌برداری از اطّلاعات، یک چهارچوب مناسب برای بررسی میزان سوییدگی نتایج حاصل از چنین مطالعاتی است. در این بخش بررسی می‌کنیم که هر روش، تحت چه شرایطی منجر به سوییدگی شده و تحت چه شرایطی سوییدگی آن قابل صرف نظر و کوچک است.

۱.۳ رتبه‌بندی با وجود سیگنال در داده‌ها

مثالی از مسئله‌ی رتبه بندی، انتخاب K مقدار i با بیشترین ϕ_i است. مثالی خاصی از این مسئله، یافتن ماکزیمم مقادیر ϕ_i می باشد. در این بخش نشان می‌دهیم وجود سیگنال (در مقابل نویز) در داده، باعث کم شدن سوییدگی در مسئله‌ی رتبه‌بندی می‌شود. اطّلاعات متقابل $I(T;\boldsymbol{\phi})$ کران دار است: $I(T;\boldsymbol{\phi}) \leq H(T) \leq \log(m)$. به صورت شهودی، هنگامی که در داده، سیگنال بزرگی از این که کدام T باید انتخاب شود در بر دارد، توزیع T از توزیع یونیفرم دور است و اطّلاعات متقابل از $\log(m)$ کمتر است. مثال زیر را در نظر بگیرید:

$$\phi_i \sim \begin{cases} \mathcal{N}(\mu, \sigma^2) & i = I^* \\ \mathcal{N}(0, \sigma^2) & i \neq I^* \end{cases} \tag{۱۲}$$

که $\mu > 0$ است. تحلیل‌گر داده قصد دارد I^* را کشف کرده و مقدار ϕ_{I^*} را گزارش کند. برای این کار $T = \arg\max_i \phi_i$. اگر $\mu = 0$ باشد، سیگنالی در داده وجود ندارد و درنتیجه تحلیل‌گر به صورت یکنواخت یکی از مقادیر $i \in \{1, 2, \ldots, m\}$ را انتخاب می‌کند، در این حالت $I(T;\phi) = H(T) = \log(m)$ است. با افزایش μ، متغیر T بر روی I^* متمرکز می‌شود و این منجر به کاهش $H(T)$ و در نتیجه کاهش سوییدگی $\mathbb{E}[\phi_T - \mu_T]$ می‌شود. در یک شبیه‌سازی، $m = 1000$ نمونه از ϕ_i ها تولید شده‌اند که همه‌ی آن‌ها به جز یکی از توزیع نرمال استاندارد تولید شده‌اند، اما یکی از آن‌ها توزیع $\phi_{I^*} \sim \mathcal{N}(\mu, 1)$ دارد. مقدار μ در بازه‌ی $[0, 4]$ تغییر داده شده است و میانگین نتیجه در ۱۰۰۰ تکرار در شکل (۱) آمده است. این نتایج، به وضوح درستی بحث فوق را تایید می‌کند. با افزایش μ کران‌بالای سوییدگی و مقدار واقعی سوییدگی کاهش یافته است.



شکل ۱: نمودار میزان سوییدگی ϕ_T و کران‌های آن بر حسب μ

۲.۳ تفکیک تقریباً مستقل داده‌های غیر i.i.d.

در این مسئله، تحلیل‌گر داده، n نمونه‌ی زمانی از یک زنجیره‌ی مارکوفی را در اختیار دارد. این تحلیل‌گر قصد دارد از این داده به همان نحوی که از داده‌های i.i.d. استفاده می‌کند، بهره ببرد. برای این کار، او داده‌ها را به سه قسمت تقسیم می‌کند: از $s_1, \ldots, s_{n_1}$ برای انتخاب تحلیل استفاده می‌کند، داده‌های $s_{n_1}, \ldots, s_{n_2}$ را دور می‌اندازد و از $s_{n_2}, \ldots, s_n$ برای تخمین تحلیل انتخاب‌شده با استفاده از بخش اول داده‌ها بهره می‌برد. اگر داده‌ها i.i.d. می‌بودند، هیچ نشت اطّلاعاتی رخ نمی‌داد و این تحلیل هیچ سوییدگی‌ای ایجاد نمی‌کرد. اما در این حالت که در دیتا روابط زمانی وجود دارد، این گزاره صادق نیست. از نظر شهودی انتظار داریم که اگر $n_2 - n_1$ به قدر کافی بزرگ باشد، سوییدگی تحلیل به قدر دلخواه کم شود.

فرض کنید زنجیره‌ی مارکوف ایستان بوده و توزیع ایستان آن را π بنامید. همچنین فرض کنید:

$$\forall \tau \in \mathbb{N} \quad \max_s D\Big(\mathbf{P}(s_\tau = .|s_1 = s)||\pi\Big) \le c_0 e^{-c_1 \tau} \tag{۱۳}$$

یعنی توزیع به شرط حالت اولیه نیز به توزیع ایستان میل می‌کند. از آن‌جایی که این فرآیند، یک فرآیند ایستان است، داریم $\mathbf{P}(s_t = s) = \pi$. با این فرض و با کمک نامساوی (۱۳) می‌توان نوشت:

$$I(s_{t+\tau}; s_t) = \sum_s \mathbf{P}(s_t = s) D\Big(\mathbf{P}\big(s_{t+\tau} = .|s_t = s\big)||\mathbf{P}(s_t = .)\Big) \le c_0 e^{-c_1 \tau}$$

با توجه به نامساوی پردازش اطّلاعات، داریم:

$$I(T; \phi) \le I(s_1, \ldots, s_{n_1}; s_{n_2+1}, \ldots, s_n) \le I(s_{n_1}; s_{n_2+1}) \le c_0 e^{-c_1(n_2 - n_1)} \tag{۱۴}$$

با ترکیب این کران با قضیه‌ی (۱.۲)، شهود ما اثبات می‌شود یعنی هرچقدر $n_2 - n_1$ بزرگتر باشد، سوییدگی کمتر می‌شود.

۴ محدود کردن سوییدگی با تصادفی‌سازی

در این بخش، نشان می‌دهیم که حتی اگر فرآیند انتخاب T، فرآیندی پیچیده بوده و یا ناشناخته باشد، می‌توان با افزودن نویز در مرحله‌ی انتخاب، تا حد زیادی سوییدگی را کم کرد. روش‌های مبتنی بر تصادفی سازی، تاکنون به شدّت در ادبیات یادگیری ماشین و آمار تکرار شده‌اند. در این بخش قصد داریم به تحلیل این روش‌ها بر اساس معیار «اطّلاعات بد» معرفی شده در فصل‌ها بپردازیم.

۱.۴ رگولاریزیشن با انتخاب تصادفی

فرض کنید m تحلیل مختلف $\phi_1, \phi_2, \ldots, \phi_m$ در اختیار داریم. قصد داریم تحلیلی را بیابیم که بزرگترین مقدار ممکن را دارد. برای این کار، یک راه اولیه، انتخاب $T = \mathrm{argmax}_i \phi_i$ است. با توجه به (۱.۲)، می‌دانیم که اگر $I(T;\phi)$ کوچک باشد، سوییدگی تخمین کم می‌شود. داریم $I(T;\phi) = H(T) - H(T|\phi)$. با توجه به این رابطه، به نظر می‌رسد که می‌توان با افزایش $H(T|\phi)$ یا به عبارتی، تصادفی‌سازی انتخاب T، میزان سوییدگی را کنترل کرد. با انتخاب T به صورت تصادفی و مستقل از داده، $H(T|\phi)$ را بیشنه می‌کند، امّا از آن‌جایی که $H(T)$ را هم زیاد می‌کند، به کمترین $I(T;\phi)$ منتج نمی‌شود. یک ایده این است که علاوه بر افزایش $H(T|\phi)$، مقدار ϕ_T را هم تا حد امکان بزرگ انتخاب کنیم. بعد از مشاهده‌ی ϕ ها، اندیس T را به صورت تصادفی و با توزیع π از $\{1, 2, \ldots, m\}$ انتخاب می‌کنیم. از آن‌جا که قصد داریم ϕ_T بزرگ باشد، منطقی است که توزیع π ای را انتخاب کنیم که جواب مسئله‌ی بهینه‌سازی زیر باشد:

$$\begin{aligned} \underset{\pi \in \mathbb{R}_+^m}{\text{maximize}} \quad & H(\pi) \\ \text{subject to} \quad & \sum_{i=1}^{k} \pi_i \phi_i \geq b \text{ and } \sum_{i=1}^{k} \pi_i = 1. \end{aligned}$$

شرط $\sum_{i=1}^{k} \pi_i \phi_i \geq b$ تضمین می‌کند که توزیع انتخاب‌شده به نحوی باشد که به صورت متوسط، مقدار ϕ_T بزرگ باشد. این یک مسئله‌ی کلاسیک در تئوری اطّلاعات است و می‌دانیم پاسخ آن، توزیع گیبس است یعنی توزیع $\pi^* = Ae^{\beta\phi_i}$. در این توزیع، β به نحوی انتخاب شده است که شرط $\sum_{i=1}^{k} \pi_i \phi_i = b$ برقرار باشد.

اگر دو دیتاست در نظر بگیریم که تنها تفاوت کوچکی با یکدیگر دارند، ممکن است $T = \mathrm{argmax}_i \phi_i$ در این دو دیتاست تفاوت بزرگی داشته باشد، اما روش تصادفی فوق، احتمالاً جواب‌های نزدیکی را در این دو دیتاست باز خواهد گرداند.

به کمک یک شبیه‌سازی، نشان می‌دهیم این روش از سوییدگی کم می‌کند. فرض کنید دو دسته ϕ_i داریم. برای تحلیل‌های $1 \leq i \leq N_1$ داریم $\mu_i = \mu > 0$ و برای تحلیل‌های $N_1 + 1 \leq i \leq m$ نیز $\mu_i = 0$ است. در شبیه‌سازی گرفته‌ایم $N_1 = 1000$ و $m - N_1 = 100000$، یعنی تعداد تحلیل‌ها با میانگین صفر بسیار بیشتر از تعداد تحلیل‌های با میانگین ناصفر است. مقدار μ را نیز در بازه‌ی 1 تا 5 تغییر می‌دهیم. در این تحلیل به جای گزارش کردم ماکزیمم، K تحلیل با مقدار بزرگتر را گزارش می‌کنیم. در این شبیه‌سازی به دو کمیت زیر علاقه‌مندیم:

- **سوییدگی:** $\frac{1}{K}\sum_{i=1}^{K} \phi_{T_i} - \mu_{T_i}$
- **دقّت:** $\frac{|\{T_i : T_i \leq N_1\}|}{K}$

نتایج این شبیه‌سازی در نمودار (۲) آمده است. در این نمودار دیده می‌شود که در رژیم با μ کوچک، مطابق انتظار هر دو روش دقّت کمی دارند. در رژیم با μ بزرگ، هر دو روش دقتی نزدیک به ۱ دارند ولی روش تصادفی Max Entropy سوییدگی کمتری را تجربه کرده است. در رژیم $1 < \mu < 4$ روش Max Entropy خطای بیشتر و در عین حال، سوییدگی کمتری دارد.

تذکر ۱.۴. توجه نمایید که الگوریتم فوق، لزوماً منجر به کم شدن سوییدگی نمی‌شود، زیرا در عین‌حالی که $H(T|\phi)$ را افزایش می‌دهد، مقدار $H(T)$ را نیز زیاد می‌کند.

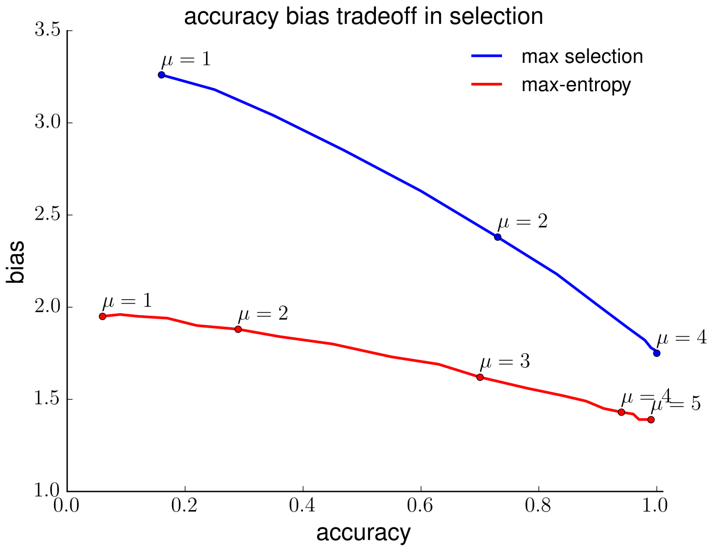
۲.۴ تصادفی‌سازی در تحلیل‌های چند مرحله‌ای

در ابتدا مدلی از تحلیل داده‌ی تطبیقی ارائه می‌کنیم در آن، تحلیل‌گر داده با انجام آنالیزهای پی‌درپی و استفاده از نتایج آنالیزهای قبلی، سعی می‌کند به شناختی از داده برسد.

- در قدم اول، تحلیل‌گر تابع ϕ_{T_1} را انتخاب می‌کند. نتیجه‌ی این تحلیل $Y_{T_1} \in \mathbb{R}$ است.
- در تکرار k ام، تحلیل‌گر تابع ϕ_{T_k} را انتخاب می‌کند و در این انتخاب خود، از متغیرهای

  $$\{Y_{T_1}, Y_{T_2}, \ldots, Y_{T_k-1}, T_1, \ldots, T_{k-1}\}$$

  استفاده می‌کند.



شکل ۲: نتایج شبیه‌سازی با تغییر μ با دو روش تصادفی این بخش و روش انتخاب ماکزیمم‌ها

ما در حالت کلّی اجازه می‌دهیم که نتیجه‌ی آنالیز kام، Y_{T_k} با ϕ_{T_k} متفاوت باشد. به عنوان مثال، ممکن است این نتیجه به صورت $Y_{T_k} = \phi_{T_k} + \text{noise}$ باشد. این سناریو زمانی رخ می‌دهد که تحلیل‌گر داده برای کاستن سوییدگی، به نتایج نویز اضافه کرده و از نتیجه‌ی نویزی در مراحل بعد تحلیل خود استفاده نماید.

در این مدل، بین متغیرهای مسئله، زنجیره‌ی ماکوف زیر برقرار است:

$$T_{k+1} - H_k = \{T_1, Y_{T_1}, T_2, Y_{T_2}, \dots, T_k, Y_{T_k}\} - D - \boldsymbol{\phi}$$

با نامساوی پردازش اطّلاعات، می‌توان گفت که $I(T_{k+1}) \leq I(H_k;\phi)$ است، بنابراین فرآیندی که در آن اطّلاعات متقابل سابقه و فیدبک سیستم (H_k) و نتایج آنالیزها $\boldsymbol{\phi}$ کنترل شده باشد، بایاس کمی دارد.

لم ۱.۴ با تعریف $H_k = \{T_1, Y_{T_1}, T_2, Y_{T_2}, \dots, T_k, Y_{T_k}\}$ داریم:

$$I(T_{k+1};\boldsymbol{\phi}) \leq I(H_k;\boldsymbol{\phi}) = \sum_{i=1}^{k} I(Y_{T_i};\phi_{T_i}|H_{i-1}, T_i)$$

از لم (۱.۴) می‌توان دریافت که با محدودکردن $I(Y_{T_i};\phi_{T_i}|H_{i-1}, T_i)$ در هر مرحله، می‌توان $I(T_{k+1};\boldsymbol{\phi})$ را کنترل کرد. در نتیجه انگار یک «بودجه‌ی اطّلاعات استفاده‌شده» مانند I_b داریم که در هر مرحله، تحلیل‌گر برای انتخاب اندیس فعلی بر اساس اندیس‌های قبلی از آن خرج می‌کند. در نتیجه فرآیند را می‌توان تا جایی ادامه داد که بودجه تمام شود.

اگر بتوانیم در هر مرحله، $I(Y_{T_i};\phi_{T_i}|H_{i-1}, T_i)$ را کم کنیم، می‌توان تعداد مراحل بیشتری را پیمود. یک راه برای دستیابی به این هدف، اضافه‌کردن نویز گاوسی است. فرض کنید $\phi_i \sim \mathcal{N}(\mu_i, \frac{\sigma^2}{n})$ و برای هر k، $(\phi_1, \cdots, \phi_k)$ مشترکاً گاوسی باشند. همچنین فرض کنید که در jامین مرحله، داریم $Y_{T_j} = \phi_{T_j} + W_j$ که W_jها مستقلند و داریم $W_j \sim \mathcal{N}(0, \frac{\omega_j^2}{n})$

در این حالت می‌توان $\mathbb{E}\left[|Y_{T_{k+1}} - \mu_{T_{k+1}}|\right]$ را محدود کرد.

قضیه ۱.۴ فرض کنید $\phi_i \sim \mathcal{N}(\mu_i, \frac{\sigma^2}{n})$ و برای هر k، $(\phi_1, \cdots, \phi_k)$ مشترکاً گاوسی باشند. همچنین فرض کنید که به ازای هر j، داریم $Y_{T_j} = \phi_{T_j} + W_j$ که W_jها مستقلند و داریم $W_j \sim \mathcal{N}(0, \frac{\sigma^2\sqrt{j}}{n})$. در این صورت برای هر $k \in \mathcal{N}$ داریم:

$$\mathbb{E}\left[|Y_{T_{k+1}} - \mu_{T_{k+1}}|\right] \leq c\left(\frac{\sigma k^{\frac{1}{4}}}{n^{\frac{1}{2}}}\right) \quad (۱۵)$$

که c یک ثابت است و مستقل از σ, k, n.

اگر دنباله‌ی $(T_1, T_2, \dots)$ به صورت غیر تطبیقی تولید می‌شدند و نویزی نیز اضافه نمی‌شد، به کران $\mathbb{E}[|Y_{T_{k+1}} - \mu_{T_{k+1}}|] \leq \frac{\sigma}{n}$ می‌رسیدیم. در مدل تطبیقی و در مراحل ابتدایی نیز می‌توانستیم به این سوییدگی برسیم، اما به مرور میزان سوییدگی افزایش پیدا می‌کرد. ضریب $k^{\frac{1}{4}}$ را می‌توان به عنوان هزینه‌ی پردازش تطبیقی در نظر گرفت. در این‌جا ذکر این نکته جالب است که می‌توان مثال‌هایی تطبیقی یافت که بدون افزودن نویز، در آن‌ها سوییدگی از حالت غیر تطبیقی بیشتر است، یعنی $\mathbb{E}[|Y_{T_{k+1}} - \mu_{T_{k+1}}|] = \Omega(\sigma\frac{k}{n})$.

۵ پیشنهاد برای کارهای آتی

برای انجام کارهای آتی چندین مسیر امیدوارکننده به نظر می‌رسد که در این بخش، به صورت اجمالی به آن‌ها اشاره می‌کنیم.

- در مدل‌سازی این مقاله، تحلیل‌گر داده همواره تنها یک تابع را بر می‌گزینند و مقدار آن را گزارش می‌کند. در کاربردهای تجربی، بسیار پیش می‌آید که تعدادی تحلیل انتخاب شده و نتایج آن گزارش شوند. به عنوان مثال، ممکن است این تعداد با توجه به داده، تعداد تحلیل‌های جالب مشخص شود. این انتخاب تصادفی تعداد داده‌های گزارش شده نیز می‌تواند باعث نشر «اطّلاعات بد» از دیتاست شده و باعث افزایش سوییدگی گردد. تعمیم چهارچوب مطرح شده در این مقاله برای حالتی که تحلیل‌گر تعدادی تحلیل را گزارش می‌کند و خود این تعداد نیز یک متغیر تصادفی است، از اهمیت تجربی و تئوری بالایی برخوردار است.

- در این مقاله و در مدل‌سازی تحلیل تطبیقی داده‌ها، همواره فرض بر این است که دیتاست در طول زمان ثابت است. این فرض در بسیاری از کاربردها از واقعیّت به دور است. فرض بهتر این است که خود دیتاست نیز ماتریسی تصادفی در نظر گرفته‌شود که در طول زمان در حال تغییر است. مبه عنوان مثال، این ماتریس می‌توان یک فرآیند تصادفی مارکوف را تشکیل دهد. مدل‌سازی مسئله در این شرایط و با اعمال محدودیت‌هایی از تغییرات ممکن در دیتاست، می‌تواند به نتایج بسیار با اهمیتی منجر شود. در این مدل جدید می‌توان به بررسی این موضوع پرداخت که استفاده داده‌های دیتاست در بخش از زمان‌ها برای انتخاب تحلیل و استفاده از زمان‌هایی دیگر برای تخمین آن، به چه میزان می‌تواند ایجاد سوییدگی کند.

- در مدل تحلیل تطبیقی داده، به خصوص در مثال متغیرهای گاوسی، برای ارائه‌ی کران سوییدگی، توزیع نویز به نحوی انتخاب‌شده که محاسبات ساده گردند. پیداکردن توزیع‌های نویزی که منجر به کوچکترین سوییدگی شوند می‌تواند بسیار جالب و کاربردی باشد.

- در تئوری یادگیری ماشیمن، الگوریتم‌های آنلاین متعددی وجود دارند که به صورت «تطبیقی» به پردازش داده‌ها می‌پردازند. چهارچوب فعلی می‌تواند به عنوان ابزاری نیرومند برای تحلیل بایاس‌های چنین الگوریتم‌های به کار رود.

مراجع

[1] D. Russo and J. Zou, “How much does your data exploration overfit? controlling bias via information usage,” *IEEE Transactions on Information Theory*, vol.66, no.1, pp.302–323, 2019.

[2] O. Bousquet and A. Elisseeff, “Stability and generalization,” *Journal of Machine Learning Research*, vol.2, no.Mar, pp.499–526, 2002.

[3] T. Poggio, R. Rifkin, S. Mukherjee, and P. Niyogi, “General conditions for predictivity in learning theory,” *Nature*, vol.428, no.6981, pp.419–422, 2004.

[4] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan, “Learnability, stability and uniform convergence,” *Journal of Machine Learning Research*, vol.11, no.Oct, pp.2635–2670, 2010.

[5] D. McAllester, “A pac-Bayesian tutorial with a dropout bound,” *arXiv preprint arXiv:1307.2118*, 2013.

[6] M. Mohri, A. Rostamizadeh, and A. Talwalkar, "Foundations of machine learning," 2018.

آ پیش‌نیازهای موردنیاز جهت فهم اثبات‌های قضایا

۱.آ متغیّرهای تصادفی زیر–گاوسی

تعریف آ.۱. متغیّر تصادفی X با میانگین $\mu = \mathbb{E}[X]$ را زیر–گاوسی می‌نامیم هرگاه عدد مثبتی مانند σ وجود داشته باشد، به قسمی که برای هر $\lambda \in \mathbb{R}$ داشته باشیم:

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\frac{\sigma^2\lambda^2}{2}}$$

ثابت σ را پارامتر این متغیّر تصادفی می‌نامیم. به عنوان مثال، یک متغیّر تصادفی گاوسی با واریانس σ^2، خود یک متغیّر تصادفی زیر–گاوسی با پارامتر σ است. هم‌چنین تعداد زیادی از متغیّرهای تصادفی غیر گاوسی، زیر–گاوسی هستند.

قضیه آ.۱. برای هر متغیّر تصادفی زیر–گاوسی X با متوسّط $\mu = \mathbb{E}[X]$ و پارامتر σ داریم:

$$\mathbb{P}[|X-\mu| \geq t] \leq 2e^{-\frac{t^2}{2\sigma^2}} \tag{۱۶}$$

اثبات. از نامساوی مارکف می‌دانیم:

$$\mathbb{P}[X-\mu \geq t] = \mathbb{P}[e^{\lambda(X-\mu)} \geq e^{\lambda t}] \leq \frac{\mathbb{E}\left[e^{\lambda(X-\mu)}\right]}{e^{\lambda t}}$$

حال با توجّه به تعریف متغیّرهای تصادفی زیر–گاوسی داریم:

$$\mathbb{P}[X-\mu \geq t] \leq \frac{\mathbb{E}\left[e^{\lambda(X-\mu)}\right]}{e^{\lambda t}} \leq \exp(\frac{\sigma^2\lambda^2}{2} - \lambda t)$$

نامساوی بالا به ازای هر $\lambda \in \mathbb{R}$ برقرار است، من‌جمله λای که طرف راست را کمینه کند. در نتیجه:

$$\mathbb{P}[X-\mu \geq t] \leq \inf_{\lambda}\left\{\exp(\frac{\sigma^2\lambda^2}{2} - \lambda t)\right\} = e^{-\frac{t^2}{2\sigma^2}} \tag{۱۷}$$

همچنین اگر متغیّر تصادفی X، زیر–گاوسی باشد، $-X$ هم زیر–گاوسی است و به طور مشابه، داریم:

$$\mathbb{P}[-X+\mu \geq t] = \mathbb{P}[X-\mu \leq -t] \leq e^{-\frac{t^2}{2\sigma^2}}$$

و می‌توان نوشت:

$$\mathbb{P}[|X-\mu| \geq t] \leq \mathbb{P}[X-\mu \geq t] + \mathbb{P}[X-\mu \leq -t] \leq 2e^{-\frac{t^2}{2\sigma^2}}$$

□

متغیّرهای تصادفی زیر–گاوسی خواصّ گوناگونی دارند، تعدادی از این خواص را در قضایای بعدی مشاهده می‌کنیم.

قضیه آ.۲. فرض کنید X یک متغیّر تصادفی زیر–گاوسی با امید ریاضی $\mathbb{E}[X] = 0$ باشد، در این صورت اگر متغیّر تصادفی Z را به صورت $Z \sim \mathcal{N}(0, 2\sigma^2)$ در نظر بگیریم، داریم:

$$\mathbb{P}[|X| \geq s] \leq \sqrt{8}e\mathbb{P}[|Z| \geq s] \qquad \forall s \geq 0 \tag{۱۸}$$

اثبات. از قضیه‌ی آ.۱ داریم:

$$\mathbb{P}[X \geq t] \leq e^{-\frac{t^2}{2\sigma^2}} \qquad \forall t \geq 0$$

و از طرف دیگر، با توجّه به کران Mills ratio برای توزیع‌های گاوسی داریم:

$$\mathbb{P}[Z \geq t] \geq \left(\frac{\sqrt{2}\sigma}{t} - \frac{(\sqrt{2}\sigma)^3}{t^3} \right) e^{-\frac{t^2}{4\sigma^2}}$$

حال دو حالت زیر را در نظر می‌گیریم: حالتی که $t \in [0, 2\sigma]$ باشد. در این حالت، داریم:

$$\mathbb{P}[Z \geq t] \geq \mathbb{P}[Z \geq 2\sigma] = \left(\frac{1}{\sqrt{2}} - \frac{1}{2\sqrt{2}} \right) e^{-1} = \frac{1}{\sqrt{8e}}$$

و از آن‌جا که $\mathbb{P}[X \geq t] \leq 1$، داریم:

$$\frac{\mathbb{P}[X \geq t]}{\mathbb{P}[Z \geq t]} \leq \sqrt{8e}$$

حالتی که $t > 2\sigma$ باشد. در این حالت اگر کران Mills ratio را با کران به دست آمده در قضیه‌ی آ.۱ ترکیب کنیم و تعریف کنیم $s = \frac{t}{\sigma}$، داریم:

$$\begin{aligned}\sup_{t > 2\sigma} \frac{\mathbb{P}[X \geq t]}{\mathbb{P}[Z \geq t]} &\leq \sup_{s > 2} \frac{e^{-\frac{s^2}{4}}}{\frac{\sqrt{2}}{s} - \frac{2\sqrt{2}}{s^3}} \\ &\leq \sup_{s > 2} s^3 e^{-\frac{s^2}{4}} \\ &\leq \sqrt{8e}\end{aligned}$$

پس در هر دو حالت نامساوی (۱۸) برقرار است. □

قضیه آ.۳. فرض کنید X یک متغیّر تصادفی با امید ریاضی $\mathbb{E}[X] = 0$ باشد و بتوانیم عدد ثابتی مانند c و یک متغیّر تصادفی $Z \sim \mathcal{N}(0, \tau^2)$ بیابیم، به قسمی که

$$\mathbb{P}[|X| \geq s] \leq c\mathbb{P}[|Z| \geq s] \qquad \forall s \geq 0 \tag{۱۹}$$

در این صورت ثابتی مانند $\theta \geq 0$ وجود دارد، به گونه‌ای که:

$$\mathbb{E}\left[X^{2k}\right] \leq \frac{(2k)!}{2^k k!} \theta^{2k} \qquad \forall k \in \mathbb{N} \tag{۲۰}$$

اثبات. اگر $Z \sim \mathcal{N}(0, \tau^2)$ و X یک متغیّر تصادفی باشد که در رابطه‌ی $\mathbb{P}[|X| \geq s] \leq c\mathbb{P}[|Z| \geq s]$ به ازای هر $s \geq 0$ صدق کند، با توجّه به این‌که X^{2k} یک متغیّر تصادفی نامنفی است، می‌توان نوشت:

$$\begin{aligned}\mathbb{E}\left[X^{2k}\right] &= \int_0^{+\infty} \mathbb{P}[X^{2k} > s] ds \\ &= \int_0^{+\infty} \mathbb{P}[|X| > s^{\frac{1}{2k}}] ds \\ &\leq c \int_0^{+\infty} \mathbb{P}[|Z| > s^{\frac{1}{2k}}] ds \\ &= c\mathbb{E}\left[Z^{2k}\right]\end{aligned}$$

و از آن‌جا که متغیّر تصادفی Z، گاوسی است، می‌دانیم:

$$\mathbb{E}\left[Z^{2k}\right] = \frac{(2k)!}{2^k k!}\tau^{2k} \qquad \forall k \in \mathcal{N}$$

در نتیجه:

$$\begin{aligned}\mathbb{E}\left[X^{2k}\right] &\leq c\mathbb{E}\left[Z^{2k}\right] \\ &= c\frac{(2k)!}{2^k k!}\tau^{2k} \\ &\leq \frac{(2k)!}{2^k k!}(c\tau)^{2k}\end{aligned}$$

در نتیجه اگر قرار دهیم $\theta = c\tau$ به حکم می‌رسیم. $\square$

قضیه آ.۴. فرض کنید X یک متغیّر تصادفی با امید ریاضی $\mathbb{E}[X] = 0$ باشد و ثابتی مانند $\theta \geq 0$ وجود داشته باشد، به قسمی که:

$$\mathbb{E}\left[X^{2k}\right] \leq \frac{(2k)!}{2^k k!}\theta^{2k} \qquad \forall k \in \mathbb{N} \tag{۲۱}$$

در این صورت متغیّر تصادفی X، زیرگاوسی با پارامتر $\sigma = \theta\sqrt{2}$ است.

اثبات. به ازای هر $\lambda \in \mathbb{R}$ داریم:

$$\mathbb{E}\left[e^{\lambda X}\right] \leq 1 + \sum_{k=2}^{\infty} \frac{|\lambda|^k \mathbb{E}\left[|X|^k\right]}{k!}$$

(جمله‌ی متناظر با $k = 1$ به دلیل این‌که $\mathbb{E}[X] = 0$ حذف شده است) اگر X حول صفر تقارن داشته باشد، جملات دارای k فرد از بین می‌روند و داریم:

$$\mathbb{E}\left[e^{\lambda X}\right] \leq 1 + \sum_{k=1}^{\infty} \frac{\lambda^(2k)\mathbb{E}\left[|X|^k\right]}{(2k)!}\frac{(2k)!\theta^{2k}}{2^k k!} = e^{\frac{\lambda^2\theta^2}{2}}$$

که نتیجه می‌دهد X یک متغیّر تصادفی زیر-گاوسی با پارامتر θ است.
اگر X متقارن نباشد، می‌توانیم یک کران بالا برای جملات مربوط به kهای فرد بیابیم:

$$\begin{aligned}\mathbb{E}\left[|\lambda X|^{2k+1}\right] &\leq \sqrt{\mathbb{E}\left[|\lambda X|^{2k}\right]\mathbb{E}\left[|\lambda X|^{2k+2}\right]} \\ &\leq \frac{1}{2}\left(\lambda^{2k}\mathbb{E}\left[X^{2k}\right] + \lambda^{2k+2}\mathbb{E}\left[X^{2k+2}\right]\right)\end{aligned}$$

(نامساوی اوّل از نامساوی کوشی-شوارتز نتیجه می‌شود و نامساوی دوم از نامساوی میانگین حسابی-هندسی.) در نتیجه داریم:

$$\begin{aligned}\mathbb{E}\left[|\lambda X|^{2k+1}\right] &\leq 1 + \left(\frac{1}{2} + \frac{1}{2\times 3!}\right)\lambda^2\mathbb{E}\left[X^2\right] + \\ &+ \sum_{k=2}^{\infty}\left(\frac{1}{(2k)!} + \frac{1}{2}\left[\frac{1}{(2k-1)!} + \frac{1}{(2k+1)!}\right]\right)\lambda^{2k}\mathbb{E}\left[X^{2k}\right] \\ &\leq \sum_{k=0}^{\infty} 2^k \frac{\lambda^{2k}\mathbb{E}\left[X^{2k}\right]}{(2k)!} \\ &\leq \sum_{k=0}^{\infty} 2^k \frac{\lambda^{2k}}{(2k)!}\frac{(2k)!\theta^{2k}}{2^k k!} \\ &= \exp\left(\frac{(\sqrt{2}\lambda\theta)^2}{2}\right)\end{aligned}$$

که نتیجه می‌دهد X یک متغیّر تصادفی زیر–گاوسی با پارامتر $\theta\sqrt{2}$ است. $\square$

قضیه آ.۵. فرض کنید X یک متغیّر تصادفی با امید ریاضی $\mathbb{E}[X]=0$ باشد و بتوانیم عدد ثابتی مانند c و یک متغیّر تصادفی $Z\sim\mathcal{N}(0,\tau^2)$ بیابیم، به قسمی که

$$\mathbb{P}[|X|\geq s]\leq c\mathbb{P}[|Z|\geq s] \qquad \forall s\geq 0 \tag{۲۲}$$

در این صورت X یک متغیّر تصادفی زیر–گاوسی با پارامتر $\sqrt{2}c\tau$ است.

اثبات. با توجّه به اینکه

$$\mathbb{P}[|X|\geq s]\leq c\mathbb{P}[|Z|\geq s] \qquad \forall s\geq 0$$

با استفاده از قضیه‌ی ۳.آ داریم:

$$\mathbb{E}\left[X^{2k}\right]\leq\frac{(2k)!}{2^k k!}\theta^{2k} \qquad \forall k\in\mathbb{N}$$

که $\theta=c\tau$. حال با توجّه به قضیه‌ی ۴.آ می‌توانیم نتیجه بگیریم که متغیّر تصادفی X، یک متغیّر تصادفی زیر–گاوسی با پارامتر $\sigma=\sqrt{2}\theta=\sqrt{2}c\tau$ است. $\square$

قضیه آ.۶. فرض کنید X یک متغیّر تصادفی زیر–گاوسی با امید ریاضی $\mathbb{E}[X]=0$ و پارامتر σ باشد، در این صورت:

$$\mathbb{E}\left[e^{\frac{sX^2}{2\sigma^2}}\right]\leq\frac{1}{\sqrt{1-s}} \qquad \forall s\in[0,1) \tag{۲۳}$$

اثبات. درستی حکم برای $s=0$ واضح است. برای $s\in(0,1)$، از آن‌جا که X یک متغیّر تصادفی زیر–گاوسی است، داریم:

$$\mathbb{E}\left[e^{\lambda X}\right]\leq e^{\frac{\lambda^2\sigma^2}{2}}$$

دو طرف نامساوی فوق را در $e^{-\frac{\lambda^2\sigma^2}{2s}}$ ضرب می‌کنیم:

$$\mathbb{E}\left[e^{\lambda X-\frac{\lambda^2\sigma^2}{2s}}\right]\leq e^{\frac{\lambda^2\sigma^2(s-1)}{2s}}$$

از آن‌جا که رابطه‌ی اخیر به ازای هر $\lambda\in\mathbb{R}$ برقرار است، می‌توانیم از دوطرف آن روی λ انتگرال بگیریم. انتگرال سمت راست نامساوی عبارت است از:

$$\int_{-\infty}^{\infty}\exp\left(\frac{\lambda^2\sigma^2(s-1)}{2s}\right)d\lambda=\frac{1}{\sigma}\sqrt{\frac{2\pi s}{1-s}}$$

و انتگرال سمت چپ نامساوی:

$$\begin{aligned}\int_{-\infty}^{\infty}\mathbb{E}\left[e^{\lambda X-\frac{\lambda^2\sigma^2}{2s}}\right]d\lambda&=\mathbb{E}\left[\int_{-\infty}^{\infty}e^{\lambda X-\frac{\lambda^2\sigma^2}{2s}}d\lambda\right]\\&=\mathbb{E}\left[\frac{\sqrt{2\pi s}}{\sigma}e^{\frac{sX^2}{2\sigma^2}}\right]\\&=\frac{\sqrt{2\pi s}}{\sigma}\mathbb{E}\left[e^{\frac{sX^2}{2\sigma^2}}\right]\end{aligned}$$

در نتیجه:

$$\begin{aligned}\mathbb{E}\left[e^{\frac{sX^2}{2\sigma^2}}\right] &= \frac{\sigma}{\sqrt{2\pi s}}\int_{-\infty}^{\infty}\mathbb{E}\left[e^{\lambda X-\frac{\lambda^2\sigma^2}{2s}}\right]d\lambda \\ &\le \frac{\sigma}{\sqrt{2\pi s}}\int_{-\infty}^{\infty}\exp\left(\frac{\lambda^2\sigma^2(s-1)}{2s}\right)d\lambda \\ &= \frac{\sigma}{\sqrt{2\pi s}}\frac{1}{\sigma}\sqrt{\frac{2\pi s}{1-s}} \\ &= \frac{1}{\sqrt{1-s}}\end{aligned}$$

□

۲.آ متغیّرهای تصادفی زیر-نمایی

تعریف ۲.آ. متغیّر تصادفی X با امید ریاضی $\mu = \mathbb{E}[X]$ را زیر-نمایی می‌نامیم اگر پارامترهای نامنفی (ν, α) وجود داشته باشند، به قسمی که برای هر λ که $|\lambda| < \frac{1}{\alpha}$ داشته باشیم:

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \le e^{\frac{\nu^2\lambda^2}{2}}$$

از تعریف متغیّرهای تصادفی زیر-گاوسی و زیر-نمایی می‌توان نتیجه گرفت که هر متغیّر تصادفی زیر-گاوسی با پارامتر σ زیر-نمایی هم هست با پارامترهای $\nu = \sigma$ و $\alpha = 0$، ولی عکس آن الزاماً درست نیست.

می‌توان قضیه‌ای مشابه قضیه‌ی ۱.آ برای متغیّرهای تصادفی زیر-نمایی بیان کرد:

قضیه ۷.آ. برای هر متغیّر تصادفی زیر-نمایی با متوسّط $\mu = \mathbb{E}[X]$ و پارامترهای ν, α داریم:

$$\mathbb{P}[X-\mu \ge t] \le \begin{cases} e^{-\frac{t^2}{2\nu^2}} & 0 \le t \le \frac{\nu^2}{\alpha} \\ e^{-\frac{t}{2\alpha}} & t > \frac{\nu^2}{\alpha} \end{cases} \tag{۲۴}$$

اثبات. از نامساوی مارکف می‌دانیم:

$$\mathbb{P}[X-\mu \ge t] = \mathbb{P}[e^{\lambda(X-\mu)} \ge e^{\lambda t}] \le \frac{\mathbb{E}\left[e^{\lambda(X-\mu)}\right]}{e^{\lambda t}}$$

و در نتیجه با استفاده از تعریف متغیّرهای تصادفی زیر-نمایی می‌توان نوشت:

$$\mathbb{P}[X-\mu \ge t] \le \frac{\mathbb{E}\left[e^{\lambda(X-\mu)}\right]}{e^{\lambda t}} \le \exp(\frac{\nu^2\lambda^2}{2} - \lambda t) \qquad \forall\lambda \in [0, \frac{1}{\alpha})$$

حال می‌خواهیم مقدار کمینه‌ی عبارت $g(\lambda, t) = \frac{\nu^2\lambda^2}{2} - \lambda t$ را محاسبه کنیم. مقدار کمینه‌ی این عبارت در $\lambda^* = \frac{t}{\nu^2}$ رخ می‌دهد. اگر داشته باشیم $0 \le t \le \frac{\nu^2}{\alpha}$، و به تبع آن $\frac{t}{\nu^2} \in [0, \frac{1}{\alpha})$ می‌توانیم λ^* را به جای λ قرار دهیم و در این صورت داریم:

$$\mathbb{P}[X-\mu \ge t] \le e^{-\frac{t^2}{2\nu^2}}$$

اگر $t > \frac{\nu^2}{\alpha}$ باشد، از آن‌جا که تابع $g(., t)$ در بازه‌ی $[0, \lambda^*]$ به طور یکنوا کاهشی است، مقدار کمینه‌ی آن در مرز $\lambda = \frac{1}{\alpha}$ رخ می‌دهد، در نتیجه:

$$\mathbb{P}[X-\mu \ge t] \le e^{-\frac{t}{\alpha}+\frac{1}{2\alpha}\frac{\nu^2}{\alpha}} \le e^{-\frac{t}{\alpha}+\frac{1}{2\alpha}t} = e^{-\frac{t}{2\alpha}}$$

□

۳.آ بیان دیگری از فاصله‌ی KL

می‌دانیم فاصله‌ی KL به صورت زیر تعریف می‌شود:

$$D_{KL}(p||q) = \sum_{x\in\mathcal{X}} p(x)\ln(\frac{p(x)}{q(x)}) \qquad (۲۵)$$

تعریف معادلی برای فاصله‌ی KL در قضیه‌ی زیر بیان شده است:

قضیه ۸.آ. فرض کنید $p(x)$ و $q(x)$ دو توزیع روی الفبای یکسان $\mathcal{X}$ باشند. در این صورت می‌توان نوشت:

$$D_{KL}(p||q) = \sup_f \left\{\mathbb{E}_p[f(X)] - \ln \mathbb{E}_q[e^{f(X)}]\right\} \qquad (۲۶)$$

که سوپریمم روی همه‌ی توابع f گرفته می‌شود که در آن‌ها $\mathbb{E}_p[f(X)]$ و $\mathbb{E}_q[e^{f(X)}]$ خوش‌تعریف باشند.

۱.۳.آ اثبات قضیه‌ی (۸.آ)

اثبات. ابتدا فرض کنید تابع f را به این صورت تعریف کنیم:

$$f(x) = \ln(\frac{p(x)}{q(x)}).$$

در این حالت مشاهده می‌کنیم که:

$$\begin{aligned}\mathbb{E}_p[f(X)] - \ln \mathbb{E}_q[e^{f(X)}] &= \sum_{x\in\mathcal{X}} p(x)\ln(\frac{p(x)}{q(x)}) - \ln(\sum_{x\in\mathcal{X}} q(x)\frac{p(x)}{q(x)}) \\ &= \sum_{x\in\mathcal{X}} p(x)\ln(\frac{p(x)}{q(x)}) - \ln(1) \\ &= \sum_{x\in\mathcal{X}} p(x)\ln(\frac{p(x)}{q(x)}) = D_{KL}(p||q)\end{aligned}$$

در نتیجه:

$$\begin{aligned}\sup_f \left\{\mathbb{E}_p[f(X)] - \ln \mathbb{E}_q[e^{f(X)}]\right\} &\geq \left[\mathbb{E}_p[f(X)] - \ln \mathbb{E}_q[e^{f(X)}]\right]_{f(x)=\ln(\frac{p(x)}{q(x)})} \\ &= D_{KL}(p||q)\end{aligned} \qquad (۲۷)$$

از طرف دیگر، داریم:

$$\begin{aligned}\mathbb{E}_p[f(X)] - \ln \mathbb{E}_q[e^{f(X)}] &= \mathbb{E}_p[f(X)] - \mathbb{E}_p[\ln \mathbb{E}_q[e^{f(X)}]] \\ &= \mathbb{E}_p\left[f(X) - \ln \mathbb{E}_q[e^{f(X)}]\right] \\ &= \mathbb{E}_p\left[\ln \frac{e^{f(X)}}{\mathbb{E}_q[e^{f(X)}]}\right] \\ &= \sum_{x\in\mathcal{X}} p(x)\ln \frac{e^{f(x)}}{\sum_{x'\in\mathcal{X}} q(x')e^{f(x')}}\end{aligned}$$

اگر توزیع $q^{(f)}(x)$ را به صورت زیر تعریف کنیم:

$$q^{(f)}(x) = \frac{q(x)e^{f(x)}}{\sum_{x' \in \mathcal{X}} q(x')e^{f(x')}}$$

می‌توانیم بنویسیم:

$$\begin{aligned}\mathbb{E}_p[f(X)] - \ln \mathbb{E}_q[e^{f(X)}] &= \sum_{x \in \mathcal{X}} p(x) \ln \frac{e^{f(x)}}{\sum_{x' \in \mathcal{X}} q(x')e^{f(x')}} \\ &= \sum_{x \in \mathcal{X}} p(x) \ln \frac{q^{(f)}(x)}{q(x)}\end{aligned}$$

و در نتیجه:

$$\begin{aligned}D_{KL}(p||q) - \Big(\mathbb{E}_p[f(X)] - \ln \mathbb{E}_q[e^{f(X)}]\Big) &= \sum_{x \in \mathcal{X}} p(x) \ln(\frac{p(x)}{q(x)}) - \sum_{x \in \mathcal{X}} p(x) \ln(\frac{q^{(f)}(x)}{q(x)}) \\ &= \sum_{x \in \mathcal{X}} p(x) \ln(\frac{p(x)}{q^{(f)}(x)}) \\ &= D_{KL}(p||q^{(f)}) \geq 0\end{aligned}$$

در نتیجه:

$$D_{KL}(p||q) \geq \sup_f \left\{\mathbb{E}_p[f(X)] - \ln \mathbb{E}_q[e^{f(X)}]\right\} \tag{۲۸}$$

و از مقایسه‌ی (۲۷) و (۲۸) داریم:

$$D_{KL}(p||q) = \sup_f \left\{\mathbb{E}_p[f(X)] - \ln \mathbb{E}_q[e^{f(X)}]\right\}$$

□

ب اثبات قضایای بیان‌شده

ب.۱ قضایای فصل ۲

ب.۱.۱ اثبات قضیه‌ی (۱.۲)

اثبات. با توجّه به این‌که $\boldsymbol{\phi} = (\phi_1, \cdots, \phi_m)$ یک متغیّر تصادفی برداری پیوسته است و T یک متغیّر تصادفی اسکالر گسسته، اگر تحقّق‌های $\boldsymbol{\phi}$ را با $\boldsymbol{\varphi} = (\varphi_1, \cdots, \varphi_m)$ و تحقّق‌های T را با i نشان دهیم، داریم:

$$\begin{aligned}I(T;\boldsymbol{\phi}) &= \sum_{i=1}^{m} \int_{\mathbb{R}^m} f_{\boldsymbol{\phi},T}(\boldsymbol{\varphi}, i) \log \frac{f_{\boldsymbol{\phi},T}(\boldsymbol{\varphi}, i)}{p_T(i) f_{\boldsymbol{\phi}}(\boldsymbol{\varphi})} d\boldsymbol{\varphi} && (۲۹) \\ &= \sum_{i=1}^{m} \int_{\mathbb{R}^m} f_{\boldsymbol{\phi}|T}(\boldsymbol{\varphi}|i) p_T(i) \log \frac{f_{\boldsymbol{\phi}|T}(\boldsymbol{\varphi}|i)}{f_{\boldsymbol{\phi}}(\boldsymbol{\varphi})} d\boldsymbol{\varphi} && (۳۰) \\ &= \sum_{i=1}^{m} p_T(i) \int_{\mathbb{R}^m} f_{\boldsymbol{\phi}|T}(\boldsymbol{\varphi}|i) \log \frac{f_{\boldsymbol{\phi}|T}(\boldsymbol{\varphi}|i)}{f_{\boldsymbol{\phi}}(\boldsymbol{\varphi})} d\boldsymbol{\varphi} && (۳۱) \\ &= \sum_{i=1}^{m} \mathbb{P}[T=i] D_{KL}(f_{\boldsymbol{\phi}|T}||f_{\boldsymbol{\phi}}) && (۳۲)\end{aligned}$$

و اگر تعریف کنیم:

$$\boldsymbol{\phi}_{/i} = (\phi_1, \cdots, \phi_{i-1}, \phi_{i+1}, \cdots, \phi_m), \qquad \boldsymbol{\varphi}_{/i} = (\varphi_1, \cdots, \varphi_{i-1}, \varphi_{i+1}, \cdots, \varphi_m)$$

می‌توان نوشت:

$$I(T; \boldsymbol{\phi}) = \sum_{i=1}^{m} p_T(i) \int_{\mathbb{R}} \int_{\mathbb{R}^{m-1}} f_{\phi_i|T}(\varphi_i|i) f_{\boldsymbol{\phi}_{/i}|T,\phi_i}(\boldsymbol{\varphi}_{/i}|i, \varphi_i) \log \frac{f_{\phi_i|T}(\varphi_i|i) f_{\boldsymbol{\phi}_{/i}|T,\phi_i}(\boldsymbol{\varphi}_{/i}|i, \varphi_i)}{f_{\phi_i}(\varphi_i) f_{\boldsymbol{\phi}_{/i}}(\boldsymbol{\varphi}_{/i})} d\boldsymbol{\varphi}_{/i} d\varphi_i \tag{۳۳}$$

$$\begin{aligned}
&= \sum_{i=1}^{m} p_T(i) \int_{\mathbb{R}} \int_{\mathbb{R}^{m-1}} f_{\phi_i|T}(\varphi_i|i) f_{\boldsymbol{\phi}_{/i}|T,\phi_i}(\boldsymbol{\varphi}_{/i}|i, \varphi_i) \log \frac{f_{\phi_i|T}(\varphi_i|i)}{f_{\phi_i}(\varphi_i)} d\boldsymbol{\varphi}_{/i} d\varphi_i \\
&+ \sum_{i=1}^{m} p_T(i) \int_{\mathbb{R}} \int_{\mathbb{R}^{m-1}} f_{\phi_i|T}(\varphi_i|i) f_{\boldsymbol{\phi}_{/i}|T,\phi_i}(\boldsymbol{\varphi}_{/i}|i, \varphi_i) \log \frac{f_{\boldsymbol{\phi}_{/i}|T,\phi_i}(\boldsymbol{\varphi}_{/i}|i, \varphi_i)}{f_{\boldsymbol{\phi}_{/i}}(\boldsymbol{\varphi}_{/i})} d\boldsymbol{\varphi}_{/i} d\varphi_i && (۳۴) \\
&\geq \sum_{i=1}^{m} p_T(i) \int_{\mathbb{R}} f_{\phi_i|T}(\varphi_i|i) \log \frac{f_{\phi_i|T}(\varphi_i|i)}{f_{\phi_i}(\varphi_i)} d\varphi_i && (۳۵) \\
&= \sum_{i=1}^{m} \mathbb{P}[T = i] D_{KL}(f_{\phi_i|T} || f_{\phi_i}) && (۳۶)
\end{aligned}$$

حال اگر تعریف کنیم $g(\phi_i) = \lambda(\phi_i - \mu_i)$ و از قضیه‌ی (؟؟) با $p = f_{\phi_i|T}$ و $q = f_{\phi_i}$ استفاده کنیم، داریم:

$$\begin{aligned}
D_{KL}(f_{\phi_i|T} || f_{\phi_i}) &= \sup_f \left\{ \mathbb{E}_p[f(\phi_i)] - \ln \mathbb{E}_q[e^{f(\phi_i)}] \right\} && (۳۷) \\
&= \sup_f \left\{ \mathbb{E}[f(\phi_i)|T = i] - \ln \mathbb{E}[e^{f(\phi_i)}] \right\} && (۳۸) \\
&\geq \mathbb{E}[g(\phi_i)|T = i] - \ln \mathbb{E}[e^{g(\phi_i)}] && (۳۹) \\
&= \mathbb{E}[\lambda(\phi_i - \mu_i)|T = i] - \ln \mathbb{E}[e^{\lambda(\phi_i - \mu_i)}] && (۴۰)
\end{aligned}$$

و از آن‌جا که طبق فرض قضیه، متغیّر تصادفی $\phi_i - \mu_i$ یک متغیّر تصادفی زیر–گاوسی با پارامتر σ فرض شده است، با توجّه به اینکه $\mathbb{E}[\phi_i - \mu_i] = \mathbb{E}[\phi_i] - \mu_i = \mu_i - \mu_i = 0$، از تعریف (؟؟) داریم:

$$\mathbb{E}[e^{\lambda(\phi_i - \mu_i)}] \leq e^{\frac{\lambda^2 \sigma^2}{2}} \tag{۴۱}$$

و در نتیجه:

$$\begin{aligned}
D_{KL}(f_{\phi_i|T} || f_{\phi_i}) &\geq \mathbb{E}[\lambda(\phi_i - \mu_i)|T = i] - \ln \mathbb{E}[e^{\lambda(\phi_i - \mu_i)}] && (۴۲) \\
&\geq \lambda(\mathbb{E}[\phi_i|T = i] - \mu_i) - \frac{\lambda^2 \sigma^2}{2} && (۴۳)
\end{aligned}$$

اگر تعریف کنیم $\Delta_i = \mathbb{E}[\phi_i|T = i] - \mu_i$، با توجّه به اینکه (۴۲) برای همه‌ی مقادیر λ برقرار است، می‌توان نوشت:

$$\begin{aligned}
D_{KL}(f_{\phi_i|T} || f_{\phi_i}) &\geq \sup_\lambda \left\{ \lambda \Delta_i - \frac{\lambda^2 \sigma^2}{2} \right\} && (۴۴) \\
&= \left[\lambda \Delta_i - \frac{\lambda^2 \sigma^2}{2} \right]_{\lambda = \frac{\Delta_i}{\sigma^2}} = \frac{\Delta_i^2}{2\sigma^2} && (۴۵)
\end{aligned}$$

و در نتیجه با ترکیب روابط (۳۶) و (۴۵) داریم:

$$2\sigma^2 I(T;\boldsymbol{\phi}) \geq \sum_{i=1}^{m} \mathbb{P}[T=i]\Delta_i^2 = \mathbb{E}[\Delta_T^2] \quad (۴۶)$$

و در نتیجه، با توجّه به خواصّ امید ریاضی شرطی و نامساوی Jensen می‌توان نوشت:

$$\begin{aligned}
|\mathbb{E}[\phi_T - \mu_T]| &\leq \mathbb{E}[|\phi_T - \mu_T|] && (۴۷)\\
&= \mathbb{E}_T[\mathbb{E}[|\phi_i - \mu_i| \big| T=i]] && (۴۸)\\
&= \mathbb{E}[|\Delta_T|] && (۴۹)\\
&= \mathbb{E}\left[\sqrt{\Delta_T^2}\right] && (۵۰)\\
&\leq \sqrt{\mathbb{E}[\Delta_T^2]} && (۵۱)\\
&\leq \sigma\sqrt{2I(T;\boldsymbol{\phi})} && (۵۲)
\end{aligned}$$

□

ب.۱.۲ اثبات قضیه‌ی (۲.۲)

اثبات. مانند اثبات قضیه‌ی (۱.۲) می‌توان نوشت:

$$I(T;\boldsymbol{\phi}) \geq \sum_{i=1}^{m} \mathbb{P}[T=i] D_{KL}(f_{\phi_i|T}||f_{\phi_i}) \quad (۵۳)$$

همچنین، از آن‌جا که $\phi_i - \mu_i$ یک متغیّر تصادفی زیر–گاوسی با پارامتر σ_i است، مشابه اثبات قضیه‌ی (۱.۲) اگر تعریف کنیم $\Delta_i = \mathbb{E}[\phi_i|T=i] - \mu_i$، داریم:

$$\begin{aligned}
D_{KL}(f_{\phi_i|T}||f_{\phi_i}) &\geq \sup_{\lambda}\left\{\lambda\Delta_i - \frac{\lambda^2\sigma_i^2}{2}\right\} && (۵۴)\\
&= \left[\lambda\Delta_i - \frac{\lambda^2\sigma_i^2}{2}\right]_{\lambda=\frac{\Delta_i}{\sigma_i^2}} = \frac{\Delta_i^2}{2\sigma_i^2} && (۵۵)
\end{aligned}$$

در نتیجه داریم:

$$|\Delta_i| \leq \sigma_i\sqrt{2D_{KL}(f_{\phi_i|T}||f_{\phi_i})} \quad (۵۶)$$

حال می‌توان نوشت:

$$\begin{aligned}
|\mathbb{E}[\phi_T - \mu_T]| &\leq \mathbb{E}[|\phi_T - \mu_T|] && (۵۷)\\
&= \mathbb{E}_T[\mathbb{E}[|\phi_i - \mu_i| \big| T = i]] && (۵۸)\\
&= \mathbb{E}[|\Delta_T|] && (۵۹)\\
&= \sum_{i=1}^{m} |\Delta_i| \mathbb{P}[T = i] && (۶۰)\\
&\leq \sum_{i=1}^{m} \sigma_i \mathbb{P}[T = i] \sqrt{2D_{KL}(f_{\phi_i|T}||f_{\phi_i})} && (۶۱)\\
&\leq \sqrt{\sum\nolimits_{i=1}^{m} \mathbb{P}[T = i]\sigma_i^2} \sqrt{2\sum\nolimits_{i=1}^{m} \mathbb{P}[T = i] D_{KL}(f_{\phi_i|T}||f_{\phi_i})} && (۶۲)\\
&\leq \sqrt{\mathbb{E}[\sigma_T^2]} \sqrt{2I(T; \boldsymbol{\phi})} && (۶۳)
\end{aligned}$$

□

۳.۱.ب اثبات قضیه‌ی (۳.۲)

اثبات. مشابه اثبات قضیه‌ی (۱.۲) می‌توان نوشت:

$$I(T; \boldsymbol{\phi}) \geq \sum_{i=1}^{m} \mathbb{P}[T = i] D_{KL}(f_{\phi_i|T}||f_{\phi_i}) \tag{۶۴}$$

همچنین، از آن‌جا که $\phi_i - \mu_i$ یک متغیّر تصادفی زیر–نمایی با پارامترهای (σ, b) است، مشابه اثبات قضیه‌ی (۱.۲) اگر تعریف کنیم $\Delta_i = \mathbb{E}[\phi_i|T = i] - \mu_i$، داریم:

$$\begin{aligned}
D_{KL}(f_{\phi_i|T}||f_{\phi_i}) &\geq \sup_{\lambda < \frac{1}{b}} \left\{ \lambda\Delta_i - \frac{\lambda^2\sigma_i^2}{2} \right\} && (۶۵)\\
&\geq \left[\lambda\Delta_i - \frac{\lambda^2\sigma_i^2}{2} \right]_{\lambda = \frac{1}{b}} = \frac{\Delta_i}{b} - \frac{\sigma^2}{2b^2} && (۶۶)
\end{aligned}$$

در نتیجه:

$$\begin{aligned}
I(T; \boldsymbol{\phi}) &\geq \sum_{i=1}^{m} \mathbb{P}[T = i] D_{KL}(f_{\phi_i|T}||f_{\phi_i}) && (۶۷)\\
&\geq \sum_{i=1}^{m} \mathbb{P}[T = i] \left(\frac{\Delta_i}{b} - \frac{\sigma^2}{2b^2} \right) && (۶۸)\\
&= \frac{\mathbb{E}_T[\mathbb{E}[\phi_i|T = i] - \mu_i]}{b} - \frac{\sigma^2}{2b^2} && (۶۹)\\
&= \frac{\mathbb{E}[\phi_T - \mu_T]}{b} - \frac{\sigma^2}{2b^2} && (۷۰)
\end{aligned}$$

و در نتیجه:

$$\mathbb{E}[\phi_T - \mu_T] \leq bI(T; \boldsymbol{\phi}) + \frac{\sigma^2}{2b} \tag{۷۱}$$

در حالتی که $b < 1$ باشد، می‌توان $\lambda = \frac{1}{\sqrt{b}} < \frac{1}{b}$ قرار داد، در نتیجه:

$$D_{KL}(f_{\phi_i|T}||f_{\phi_i}) \geq \sup_{\lambda < \frac{1}{b}} \left\{ \lambda \Delta_i - \frac{\lambda^2 \sigma_i^2}{2} \right\} \tag{۷۲}$$

$$\geq \left[\lambda \Delta_i - \frac{\lambda^2 \sigma_i^2}{2} \right]_{\lambda = \frac{1}{\sqrt{b}}} = \frac{\Delta_i}{\sqrt{b}} - \frac{\sigma^2}{2b} \tag{۷۳}$$

و داریم:

$$I(T; \boldsymbol{\phi}) \geq \sum_{i=1}^{m} \mathbb{P}[T = i] D_{KL}(f_{\phi_i|T}||f_{\phi_i}) \tag{۷۴}$$

$$\geq \sum_{i=1}^{m} \mathbb{P}[T = i] \left(\frac{\Delta_i}{\sqrt{b}} - \frac{\sigma^2}{2b} \right) \tag{۷۵}$$

$$= \frac{\mathbb{E}[\phi_T - \mu_T]}{\sqrt{b}} - \frac{\sigma^2}{2b} \tag{۷۶}$$

و:

$$\mathbb{E}[\phi_T - \mu_T] \leq \sqrt{b} I(T; \boldsymbol{\phi}) + \frac{\sigma^2}{2\sqrt{b}} \tag{۷۷}$$

$\square$

ب.۱.۴ اثبات قضیه‌ی (۴.۲)

اثبات. تعریف می‌کنیم $U_i = \phi_i - \mu_i$. بنا به فرض می‌دانیم U_i زیر–گاوسی با پارامتر σ است. همچنین تعریف می‌کنیم $\gamma_i = \mathbb{E}[|U_i|] = \mathbb{E}[|\phi_i - \mu_i|]$ و $Y_i = |U_i| - \gamma_i$. داریم:

$$\mathbb{P}[|Y_i| \geq s] = \mathbb{P}[Y_i \geq s] + \mathbb{P}[Y_i \leq -s] \tag{۷۸}$$

$$= \mathbb{P}[|U_i| \geq s + \gamma_i] + \mathbb{P}[|U_i| \leq \gamma_i - s] \tag{۷۹}$$

از قضیه‌ی (آ.۲) می‌دانیم:

$$\mathbb{P}[|U_i| \geq s + \gamma_i] \leq \sqrt{8}e\mathbb{P}[|Z| \geq s + \gamma_i] \leq \sqrt{8}e\mathbb{P}[|Z| \geq s] \tag{۸۰}$$

که در آن، $Z \sim \mathcal{N}(0, 2\sigma^2)$. از طرف دیگر، داریم:

$$\mathbb{P}[|U_i| \leq \gamma_i - s] \leq \frac{\mathbb{P}[|Z| \geq s]}{\mathbb{P}[|Z| \geq \gamma_i]} \tag{۸۱}$$

در نتیجه:

$$\mathbb{P}[|Y_i| \geq s] \leq \left(\sqrt{8}e + \frac{1}{\mathbb{P}[|Z| \geq \gamma_i]} \right) \mathbb{P}[|Z| \geq s] \tag{۸۲}$$

در نتیجه، با توجّه به قضیه‌ی (آ.۵)، Y_i یک متغیّر تصادفی زیر–گاوسی با پارامتر $2\sigma \left(\sqrt{8}e + \frac{1}{\mathbb{P}[|Z| \geq \gamma_i]} \right)$ است. از آنجا که U_i زیر–گاوسی با پارامتر σ است، واریانس آن کمتر از σ^2 خواهد بود و در نتیجه داریم:

$$\gamma_i = \mathbb{E}[|U_i|] = \mathbb{E}[\sqrt{U_i^2}] \leq \sqrt{\mathbb{E}[U_i^2]} \leq \sigma$$

که نتیجه می‌دهد:

$$\mathbb{P}[|Z| \geq \gamma_i] > \mathbb{P}[|Z| \geq \sigma] > 0.1 \qquad (۸۳)$$

در نتیجه Y_i یک متغیّر تصادفی زیر–گاوسی با پارامتر $c\sigma$ است که $c < 36$
حال داریم:

$$\mathbb{E}[|\phi_T - \mu_T| - \gamma_T] = \mathbb{E}[Y_T] \leq c\sigma\sqrt{2I(T;\boldsymbol{\phi})} \qquad (۸۴)$$

و از آنجا که برای هر i، $\gamma_i \leq \sigma$، نتیجه می‌گیریم که $\gamma_T < \sigma$ و در نتیجه:

$$\mathbb{E}[|\phi_T - \mu_T|] \leq \sigma + c\sigma\sqrt{2I(T;\boldsymbol{\phi})} \leq \sigma + 36\sigma\sqrt{2I(T;\boldsymbol{\phi})} \qquad (۸۵)$$

$\square$

ب.۱.۵ اثبات قضیه‌ی (۵.۲)

اثبات. تعریف می‌کنیم $Y_i = \phi_i - \mu_i$ و $\gamma_i = \mathbb{E}\left[(\phi_i - \mu_i)^2\right]$. از آنجا که Y_i، زیر–گاوسی با پارامتر σ است، واریانس آن کمتر از σ^2 است و در نتیجه $\gamma_i < \sigma^2$.
از طرف دیگر، از آن‌جا که Y_i یک متغیّر تصادفی زیر–گاوسی با پارامتر σ است، با توجّه به قضیه‌ی (آ.۶) داریم:

$$\mathbb{E}\left[e^{\frac{\lambda Y_i^2}{2\sigma^2}}\right] \leq \frac{1}{\sqrt{1-\lambda}} \qquad \forall \lambda \in [0,1) \qquad (۸۶)$$

و می‌توان نوشت:

$$\mathbb{E}\left[e^{\frac{\lambda(Y_i^2-\gamma_i)}{2\sigma^2}}\right] \leq \mathbb{E}\left[e^{\frac{\lambda Y_i^2}{2\sigma^2}}\right] \leq \frac{1}{\sqrt{1-\lambda}} \leq e^{10\lambda^2} \qquad \forall \lambda \in [0,0.1) \qquad (۸۷)$$

اگر تعریف کنیم $t = \frac{\lambda}{\sigma^2}$، می‌توان نوشت:

$$\mathbb{E}\left[e^{t(Y_i^2-\gamma_i)}\right] \leq e^{10\sigma^4 t^2} \qquad \forall t \in [0, \frac{0.1}{\sigma^2}) \qquad (۸۸)$$

و در نتیجه، $Y_i^2 - \gamma_i$ یک متغیّر تصادفی زیر–نمایی با پارامترهای $(\sqrt{5}\sigma^2, 10\sigma^2)$ است. حال از قضیه‌ی (۳.۲) داریم:

$$\mathbb{E}[Y_T^2] \leq 10\sigma^2 I(T;\boldsymbol{Y}^2) + \frac{5\sigma^4}{20\sigma^2}$$

که $\boldsymbol{Y}^2 = (Y_1^2, \cdots, Y_m^2)$، در نتیجه:

$$\begin{aligned}
\mathbb{E}\left[(\phi_T - \mu_T)^2\right] &\leq \gamma_T + 10\sigma^2 I(T;\boldsymbol{Y}^2) + \frac{5\sigma^4}{20\sigma^2} && (۸۹)\\
&\leq \sigma^2 + 10\sigma^2 I(T;\boldsymbol{Y}^2) + \frac{5\sigma^4}{20\sigma^2} && (۹۰)\\
&= \sigma^2(1.25 + 10I(T;\boldsymbol{Y}^2)) && (۹۱)\\
&\leq \sigma^2(1.25 + 10I(T;\boldsymbol{\phi})) && (۹۲)
\end{aligned}$$

و نامساوی آخر از نامساوی پردازش داده‌ها نتیجه شده است. $\square$

ب.۱.۶ اثبات قضیه‌ی (۶.۲)

اثبات. کران بالا از قضیه‌ی (۵.۲) قابل اثبات است، در اینجا به کران پایین می‌پردازیم.
تعریف می‌کنیم $M = \max_i \phi_i$ و $M_{-i} = \max_{j \neq i} \phi_j$. در نتیجه $T = i$ خواهد شد، اگر و تنها اگر $M_{-i} \leq \phi_i$ باشد.
تعریف می‌کنیم $I = \{i : \mathbb{E}[M_{-i}] \geq \mu_i + 1\}$، در نتیجه می‌توان آنتروپی را بسط داد:

$$H(T) = \sum_{i \notin I} \mathbb{P}[T = i] \log \left(\frac{1}{\mathbb{P}[T = i]} \right) + \sum_{i \in I} \mathbb{P}[T = i] \log \left(\frac{1}{\mathbb{P}[T = i]} \right) \tag{۹۳}$$

می‌دانیم که M و M_{-i}، به ازای هر $i \in \{1, \cdots, m\}$، زیر-گاوسی با پارامتر ۱ هستند. در نتیجه می‌توان نوشت:

$$\mathbb{E}[e^{\lambda(M - \mathbb{E}[M])}] \leq e^{\frac{\lambda^2}{2}} \tag{۹۴}$$

$$\mathbb{E}\left[(M - \mathbb{E}[M])^2\right] \leq 1 \tag{۹۵}$$

$$\mathbb{P}[M \geq \mathbb{E}[M] + \lambda] \leq e^{\frac{-\lambda^2}{2}} \tag{۹۶}$$

همچنین اگر $X \sim (0, 1)$، برای هر $x > 0$ داریم:

$$\mathbb{P}[X > x] \geq \frac{1}{\sqrt{2\pi}} \left(\frac{x}{x^2 + 1} \right) e^{\frac{-x^2}{2}} \tag{۹۷}$$

حال سعی می‌کنیم در جمله‌ی اوّل مجموع (۹۳)، یک کران پایین برای $\mathbb{P}[T = i]$ بیابیم. چون در جمله‌ی اوّل هستیم، $\mathbb{E}[M_{-i}] < \mu_i + 1$ است.

$$\mathbb{P}[T = i] = \mathbb{P}[M_{-i} < \phi_i] \tag{۹۸}$$

$$\geq \mathbb{P}[M_{-i} < \mathbb{E}[M_{-i}] + \lambda].\mathbb{P}[\phi_i > \mathbb{E}[M_{-i}] + \lambda] \tag{۹۹}$$

$$\geq \mathbb{P}[M_{-i} < \mathbb{E}[M_{-i}] + \lambda].\mathbb{P}[\phi_i > \mu_i + 1 + \lambda] \tag{۱۰۰}$$

$$\geq \left(1 - e^{\frac{-\lambda^2}{2}}\right) \frac{1}{\sqrt{2\pi}} \left(\frac{1 + \lambda}{(1 + \lambda)^2 + 1} \right) e^{\frac{-(1+\lambda)^2}{2}} \tag{۱۰۱}$$

$$\triangleq p(\lambda) \tag{۱۰۲}$$

در نتیجه:

$$\sum_{i \notin I} \mathbb{P}[T = i] \log \left(\frac{1}{\mathbb{P}[T = i]} \right) \leq \mathbb{P}[T \notin I] \max_{i \notin I} \log \left(\frac{1}{\mathbb{P}[T = i]} \right) \tag{۱۰۳}$$

$$\leq \log \left(\frac{1}{p(1)} \right) \triangleq c_{-I} \tag{۱۰۴}$$

محاسبه‌ی مستقیم نشان می‌دهد که $c_{-I} < 5$
حال به سراغ جمله‌ی دوم می‌رویم، تعریف می‌کنیم $X = \phi_i - \mu_i \sim \mathcal{N}(0, 1)$ و $Y = M_{-i} - \mu_i$. حال می‌توان نوشت:

$$\mathbb{P}(T = i) = \int_{-\infty}^{\infty} \mathbb{P}(X > x)\mathbb{P}(Y = dx) \tag{۱۰۵}$$

$$\geq \int_{1}^{\infty} \mathbb{P}(X > x)\mathbb{P}(Y = dx) \tag{۱۰۶}$$

$$= \mathbb{P}(Y \geq 1) \int_{1}^{\infty} \mathbb{P}(X > x)\mathbb{P}(Y = dx | Y \geq 1) \tag{۱۰۷}$$

$$\geq \frac{\mathbb{P}(Y \geq 1)}{\sqrt{2\pi}} \int_{1}^{\infty} \left(\frac{x}{x^2 + 1} \right) e^{-x^2/2} \mathbb{P}(Y = dx | Y \geq 1). \tag{۱۰۸}$$

و با استفاده از نامساوی Jensen داریم:

$$\log \mathbb{P}(T=i) \geq \log(1/\sqrt{2\pi}) + \log(\mathbb{P}(Y \geq 1)) + \int_1^\infty \left(\log\left(\frac{x}{x^2+1}\right) - x^2/2\right) \times \mathbb{P}(Y=dx|Y\geq 1), \tag{۱۰۹}$$

که می‌تواند به صورت زیر نوشته شود:

$$\log\left(\frac{1}{\mathbb{P}(T=i)}\right) \leq \log(\sqrt{2\pi}) + \log\left(\frac{1}{\mathbb{P}(Y\geq 1)}\right) + \mathbb{E}\left[\log\left(\frac{Y^2+1}{Y}\right)|Y>1\right] + \frac{\mathbb{E}[Y^2|Y>1]}{2}. \tag{۱۱۰}$$

اگر $Y \geq 1$ باشد، داریم $\log((Y^2+1)/Y) \leq \log(1+Y) \leq Y \leq Y^2$، در نتیجه:

$$\log\left(\frac{1}{\mathbb{P}(T=i)}\right) \leq \log(\sqrt{2\pi}) + \log\left(\frac{1}{\mathbb{P}(Y\geq 1)}\right) + 1.5\mathbb{E}[Y^2|Y>1]. \tag{۱۱۱}$$

حال،

$$\mathbb{E}[Y^2|Y>1] \frac{\leq \mathbb{E}[Y^2]}{\mathbb{P}(Y>1)} \tag{۱۱۲}$$

$$= \frac{\left(\mathbb{E}[(Y-\mathbb{E}[Y])^2] + \mathbb{E}[Y]^2\right)}{\mathbb{P}(Y>1)}. \tag{۱۱۳}$$

از آنجا که $Y = M_{-i} - \mu_i$، واریانس Y کمتر از ۱ است. در نتیجه $\mathbb{P}(Y>1) \geq 1 - 1/\sqrt{e}$ و

$$\log\left(\frac{1}{\mathbb{P}(T=i)}\right) \leq \log(\sqrt{2\pi}) + \log\left(\frac{1}{\mathbb{P}(Y\geq 1)}\right) + \frac{1.5(1+\mathbb{E}[Y]^2)}{\mathbb{P}(Y\geq 1)} \tag{۱۱۴}$$

$$< 5 + 4\mathbb{E}[Y]^2. \tag{۱۱۵}$$

حال، اگر کران‌های به دست آمده را در کنار هم قرار دهیم، داریم:

$$H(T) = \sum_i \mathbb{P}(T=i)\log(1/\mathbb{P}(T=i)) \tag{۱۱۶}$$

$$\leq c_{-I} + 5 + 4\sum_{i\in I} \mathbb{P}(T=i)(\mathbb{E}[M_{-i}] - \mu_i)^2 \tag{۱۱۷}$$

$$\leq c_{-I} + 5 + 4\sum_{i\in I} \mathbb{P}(T=i)(\mathbb{E}[M] - \mu_i)^2 \tag{۱۱۸}$$

$$\leq c_{-I} + 5 + 4\|\mathbb{E}[M] - \mu_T\|^2 \tag{۱۱۹}$$

که در آن، $\|X\| \equiv \sqrt{\mathbb{E}[X^2]}$
حال داریم:

$$\|\mathbb{E}[M] - \phi_T\| = \mathbb{E}[(\phi_T - \mathbb{E}[\phi_T])^2] \leq 1. \tag{۱۲۰}$$

و در نتیجه:

$$\mathbb{E}[M] - \mu_T\| = \|\mathbb{E}[M] - \phi_T + \phi_T - \mu_T\| \leq 1 + \|\phi_T - \mu_T\|. \tag{۱۲۱}$$

و می‌توانیم نتیجه بگیریم:

$$\|\mathbb{E}[M]-\mu_T\|^2 \leq (1+\|\phi_T-\mu_T\|)^2 \leq 2+2\|\phi_T-\mu_T\|^2 \tag{۱۲۲}$$

از اینجا نتیجه می‌شود که:

$$H(T) \leq c_{-I}+5+8+8\|\phi_T-\mu_T\|^2 \tag{۱۲۳}$$

یا:

$$\|\phi_T-\mu_T\|^2 \geq c_1 H(T)-c_2 \tag{۱۲۴}$$

که در آن $c_1=\frac{1}{8}$ و $c_2=\frac{c_{-I}+13}{8}<2.5$. □

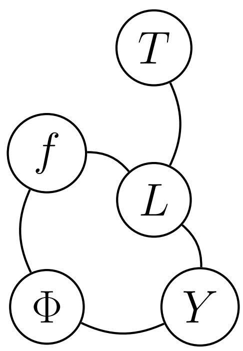
ب.۲ اثبات قضایای فصل ۳

ب.۲.۱ اثبات قضیه‌ی (پ.۲)

اثبات. بگیرید $\phi_i=\hat{L}(f_i)$ و $\mu_i=L(f_i)$. متغیر تصادفی T را نیز اندیس تصادفی‌ای در نظر بگیرید که $\hat{f}(\mathbf{x})=f_T(\mathbf{x})$.
توابع $f_1, f_2, \dots f_m$ هستند که طبقه‌بندی‌های مختلف را بر روی نمونه‌های $\mathbf{x}=(x_1,x_2,\dots,x_n)$ انجام می‌دهد (به تعریف تابع رشد مراجعه نمایید). با این تعریف‌ها، می‌توان از قضیه‌ی (۱.۲) مقاله برای بررسی مسئله‌ی طبقه‌بندی استفاده کرد. هدف، کران زدن بر روی $\mathbb{E}[\phi_T-\mu_T]$ است.
حال به بیان چند گزاره‌ی ساده در زمینه‌ی متغیرهای تصادفی زیرگاوسی می‌پردازیم. اگر $X\sim \mathrm{Bern}(p)$ متغیری تصادفی با پارامتر p باشد، آن‌گاه $X-p$ متغیری زیرگاوسی با پارامتر $\frac{1}{4}$ است. به طور مشابه، اگر $X_1,..X_n$ متغیرهای تصادفی برنولی با پارامترهای $p_1,\dots,p_n$ باشند، $n^{-1}\sum_{i=1}^{n}(X_i-p_i)$ یک متغیر تصادفی زیرگاوسی با پارامتری کمتر از $1/4\sqrt{n}$ است.
با استفاده از قضیه‌ی (۱.۲) داریم:

$$\mathbb{E}[\phi_T-\mu_T] \leq \sqrt{\frac{I(T;\boldsymbol{\phi})}{2n}}$$

شبکه‌ی مارکوفی زیر را می‌توان برای متغیرهای مسئله در نظر گرفت:



شکل ۳: شبکه‌ی مارکوفی متغیّرها در مسئله‌ی طبقه‌بندی

با توجه به این شبکه‌ی مارکوفی و با استفاده از نامساوی پردازش اطّلاعات، می‌توان نوشت:

$$I(T;\boldsymbol{\phi}) \leq I(T;\mathbf{Y}) = I(\hat{f}(\mathbf{x});\mathbf{Y})$$

از طرفی، با توجه به تعریف تابع رشد مجموعه‌ی توابع $\mathcal{F}$ داریم:

$$I(T;\boldsymbol{\phi}) \leq H(T) \leq \log(\Pi_{\mathcal{F}}(n)) \tag{۱۲۵}$$

با استفاده از لم ساور-شلاح، این عبارت را می‌توان با بُعد VC مجموعه‌ی $\mathcal{F}$ که محدود و برابر با d است، کران زد:

$$\Pi_{\mathcal{F}}(n) \leq \begin{cases} 2^n & n < d \\ (\frac{en}{d})^d & n \geq d \end{cases} \tag{۱۲۶}$$

با ترکیب (۱۲۵) و (۱۲۶) داریم: $I(\hat{f}(\mathbf{x}), \mathbf{Y}) \leq d \log_+(\frac{en}{d})$.

□

ب.۳ اثبات قضایای فصل ۴

ب.۳.۱ اثبات لم (۴.۱)

اثبات. از آن‌جایی که $H_k \perp\!\!\!\perp T_{k+1}|\phi$، با استفاده از نامساوی پردازش اطّلاعات می‌توان نوشت:

$$I(T_{k+1}; \boldsymbol{\phi}) \leq I(H_k; \boldsymbol{\phi}).$$

در نتیجه داریم:

$$I(H_k; \boldsymbol{\phi}) = \sum_{i=1}^{k} I\left((T_i, Y_{T_i}); \boldsymbol{\phi}|H_{i-1}\right).$$

لذا $I\left((T_i, Y_{T_i}); \boldsymbol{\phi}|H_{i-1}\right)$. Let $\boldsymbol{\phi}_{(-i)} = (\phi_j : j \neq i)$. با این وجود می‌توان نوشت:

$$\begin{aligned} I\left((T_i, Y_{T_i}); \boldsymbol{\phi}|H_{i-1}\right) &= I\left(T_i; \boldsymbol{\phi}|H_{i-1}\right) \\ &\quad + I\left(Y_{T_i}; \boldsymbol{\phi}|H_{i-1}, T_i\right) \\ &= I\left(Y_{T_i}; \boldsymbol{\phi}|H_{i-1}, T_i\right) \\ &= I(Y_{T_i}; \phi_{T_i}|H_{i-1}, T_i) \\ &\quad + I(Y_{T_i}; \phi_{(-T_i)}|H_{i-1}, T_i, \boldsymbol{\phi}_{T_i}) \\ &= I(Y_{T_i}; \phi_{T_i}|H_{i-1}, T_i), \end{aligned}$$

عبارت آخر از استقلال Y_{T_i} و $\phi_{(-T_i)}$ به شرط ϕ_{T_i}، Y_{T_i} به دست آمده است. به کمک این نامساوی داریم:

$$I(T_{k+1}; \phi) \leq I(H_k; \phi) = \sum_{i=1}^{k} I(Y_{T_i}; \phi_{T_i}|H_{i-1}, T_i)$$

□

ب.۳.۲ اثبات قضیه‌ی (۴.۱)

اثبات. فرض کنید $\phi_i \sim \mathcal{N}(\mu_i, \frac{\sigma^2}{n})$ و $(\phi_1, \ldots, \phi_k)$ متغیرهای مشترکاً گاوسی. اگر در مرحله‌ی j ام داشته باشیم $Y_{T_j} = \phi_{T_j} + W_j$ که $W_j \sim \mathcal{N}(0, \frac{\omega_j^2}{n})$ و متغیرهای W_j مستقل از ϕ باشند، آن‌گاه

$$\mathbb{E}[|Y_{T_{k+1}} - \mu_{T_{k+1}}|] \leq \frac{\sigma}{\sqrt{n}} + c_1 \left(\frac{\omega_{k+1}}{\sqrt{n}} + \sigma^2 \sqrt{\frac{\sum_{j=1}^{k} w_j^{-2}}{n}} \right).$$

با فرض $\omega_j = \sigma j^{\frac{1}{4}}$:

$$
\begin{aligned}
\mathbb{E}[|Y_{T_{k+1}} - \mu_{T_{k+1}}|] &\leq \mathbb{E}[|Y_{T_{k+1}} - \phi_{T_{k+1}}|] + \mathbb{E}[|\phi_{T_{k+1}} - \mu_{T_{k+1}}|] \\
&\leq \sqrt{\frac{2\omega_{k+1}}{\pi n}} + \mathbb{E}[|\phi_{T_{k+1}} - \mu_{T_{k+1}}|] \\
&\leq \sqrt{\frac{2\omega_{k+1}}{\pi n}} + \frac{\sigma}{\sqrt{n}} + c \cdot \sigma \sqrt{\frac{2I(T_{k+1};\boldsymbol{\phi})}{n}}
\end{aligned}
$$

در نامساوی دوم، از امیدریاضی متغیر تصادفی نیم-گاوسی استفاده شده است. با اعمال لم (۱.۴)، داریم:

$$
I(T_{k+1};\boldsymbol{\phi}) \leq \sum_{i=1}^{k} I(Y_{T_i};\phi_{T_i}|H_{i-1},T_i)
$$

از آن‌جایی که ϕ_i ها مشترکاً نرمال هستند، توزیع $\mathbb{P}(\phi_j|H_{i-1})$ نیز متغیری گاوسی و واریانسی کمتر از $\frac{\sigma^2}{n}$ است. همچنین به شرط H_{i+1}، متغیر T_i مستقل از $(\phi_1,\phi_2,\dots)$ و $(Y_1,Y_2,\dots)$ می‌باشد. نتیجه‌ی این بحث این است که $\phi_{T_i}|H_{i-1},T_i$ نیز دارای توزیع نرمال و واریانسی کمتر از σ^2/n است. با توجه به اطلاعات متقابل در متغیرهای گاوسی:

$$
I(Y_{T_i};\phi_{T_i}|H_{i-1},T_i) \leq \frac{\sigma^2/n}{2\omega_i^2/n} = \frac{\sigma^2}{2\omega_i^2}
$$

و بنابراین:

$$
I(T_{k+1};\boldsymbol{\phi}) \leq \left(\frac{\sigma^2}{2}\right)\sum_{i=1}^{k}\omega_i^{-2}.
$$

با جای‌گذاری این کران در عبارت مربوط به سوییدگی، می‌توان نوشت:

$$
\mathbb{E}[|Y_{T_{k+1}} - \mu_{T_{k+1}}|] \leq \sqrt{\frac{2\omega_{k+1}}{\pi n}} + \frac{\sigma}{\sqrt{n}} + c\sigma^2\sqrt{\frac{\sum_{i=1}^{k}\omega_i^{-2}}{n}},
$$

□

پ برخی دیگر از کاربردهای مسئله‌ی کاهش سوییدگی از طریق کنترل اطّلاعات استفاده شده

پ.۱ فیلتر کردن به کمک آماره‌های حاشیه‌ای

فرض کنید بعد از مشاهده‌ی دیتاست D، T انتخاب شده باشد. دیتاست D مقادیر $\phi_1,\dots,\phi_m$ را مشخص می‌کند ولی شامل اطّلاعات دیگری نیز هست. داریم:

$$
\begin{aligned}
I(T;\boldsymbol{\phi}) &= H(T) - H(T|\boldsymbol{\phi}) &(۱۲۷)\\
&\leq H(T) - I(T;D|\boldsymbol{\phi}) &(۱۲۸)\\
&= (1-\alpha)H(T) &(۱۲۹)
\end{aligned}
$$

در این معادلات، $\alpha = I(T;D|\boldsymbol{\phi})/H(T)$ است. این پارامتر، بیانگر کسر عدم قطعیتی است که علاوه بر توابع $\boldsymbol{\phi}$ در D وجود دارد. در بسیاری از مواقع، $I(T;\boldsymbol{\phi})$ خیلی کوچک‌تر از $H(T)$ است که خود کمتر از $\log(m)$ می باشد. یک مثال از این سناریو، انتخاب T بر اساس آماره‌های D می‌باشد، یک مثال از این مسئله، انتخاب ویژگی مبتنی بر واریانس است. فرض کنید n نمونه از

m ویژگی زیستی در اختیار داریم. مقدار ویژگی i ام در نمونه‌ی j ام را $X_{i,j}$ بنامید. دیتاست مورد استفاده $D = \{X_{i,j}\}$ است. تابع ϕ_i را میانگین نمونه‌ی ویژگی i ام بگیرید:

$$\phi_i = \frac{1}{n}\sum_{j=1}^{n} X_{i,j} \tag{۱۳۰}$$

ما علاقه‌مند به ویژگی‌هایی هستیم که میانگینی به طرز معنادار متفاوت با صفر داشته باشند. برخی از روش‌ها، در ابتدا یک مرحله‌ی فیلترینگ انجام می‌دهند و تنها ویژگی‌های با واریانس بزرگ را نگه می‌دارند و بقیه را حذف می‌کنند. این فیلترینگ با این استدلال انجام می‌شود که ویژگی‌های با واریانس کم، اطّلاعاتی در بر ندارند و احتمالاً خطای سیستماتیک هستند. سوال طبیعی این است که آیا این فیلترینگ منجر به بایاس می‌شود؟

مسئله را به این صورت فرمول‌بندی می‌کنیم که

$$T = \mathrm{argmax}_i \sum_{j=1}^{n}(X_{i,j}-\phi_i)^2 \tag{۱۳۱}$$

یعنی تنها یک ویژگی انتخاب شده و این ویژگی دارای بزرگترین واریانس می‌باشد. تمام نتایج امکان تعمیم به حالتی که k ویژگی با واریانس بزرگ انتخاب شده‌اند نیز می‌باشد. قضیه‌ی (۱.۲) بیان می‌کند که سوییدگی $\mathbb{E}[\phi_T - \mu_T]$ کوچک است، اگر $I(T;\boldsymbol{\phi})$ کوچک باشد.

اگر میانگین نمونه و واریانس نمونه خیلی وابسته نباشند، این شرط برقرار می‌شود. به عنوان مثال می‌دانیم که اگر نمونه‌های گاوسی و i.i.d باشند، $\phi_1,\dots,\phi_m$ از $V_1,\dots,V_m$ مستقل و $I(T;\boldsymbol{\phi}) = 0$ است، در نتیجه روش سوییدگی ایجاد نمی‌کند.

توجه داشته باشید که در این حالت $I(T;D)$ می‌تواند بسیار بزرگ باشد ولی قضیه‌ی (۱.۲) تنها به $I(T;\boldsymbol{\phi})$ وابسته است. این بدین معناست که انتخاب T تنها وابسته به بخشی از اطّلاعات دیتاست است که در توابع $\boldsymbol{\phi}$ خود را نشان نداده است.

در حالت کلّی‌تر این مسئله فرض کنید برای هر ویژگی i دو آماره‌ی حاشیه‌ای ϕ_i و ψ_i محاسبه شده‌اند. بنا بر نامساوی پردازش اطّلاعات داریم $I(T;D) \le I(\psi;\boldsymbol{\phi})$. در این جا $T = f(\psi_i)$ است و در تحلیل داده، علاقه‌مند به مقادیر ϕ_i هستیم و نتیجه‌ی تحلیل ما ϕ_T است. اگر ψ_i ها اطّلاعات زیادی از ϕ_i ها در بر نداشته‌باشند، فیلترینگ داده بر اساس ψ_i ها منجر به سوییدگی بزرگ نمی‌شود. رابطه‌ی متغیرها در این مسئله به فرم زنجیره‌ی مارکوف $T - \psi - \boldsymbol{\phi}$ مدل می‌شود. با نامساوی پردازش اطّلاعات داریم: $I(T;\boldsymbol{\phi}) \le I(\boldsymbol{\phi},\psi()$ که منتج به همان نتیجه‌ای که اگر ψ ها و ϕ ها اطّلاعات متقابل کمی داشته باشند، سوییدگی کم است می‌شود. این کران کران تیزی نیست و برای ارائه‌ی یک کران تیز، به نامساوی پردازش اطّلاعات قوی روی می آوریم.

تعریف پ.۱. متغیرهای تصادفی X و Y در نامساوی قوی پرازش اطّلاعات با پارامتر $\eta \in [0,1]$ صدق می‌کنند، اگر برای هر متغیر تصادفی U با زنجیره‌ی $U - X - Y$ داشته باشیم:

$$I(U;Y) \le \eta I(U;X)$$

η_{XY} را کوچکترین ضریبی بگیرید که نامساوی فوق برای تمام U ها برقرار باشد.

یکی از ویژگی‌های جالب ضریب η_{XY} این است که اگر $(X_1,Y_1),\dots,(X_n,Y_n)$ دنباله‌ای مستقل باشد، $\eta_{XY} = \max_i \eta_{X_iY_i}$ است. همچنین اگر برای یک متغیر تصادفی Z زنجیره‌ی $X - Y - Z$ برقرار باشد، $\eta_{XZ} \le \eta_{YZ}$ است.

مثال پ.۱. فرض کنید $D = (X_1,\dots,X_n)$ شامل n متغیر تصادفی $i.i.d$ باشد و $\psi = (X_1,\dots,X_k)$ که $k \le n$ است، یک زیرمجموعه‌ی nتایی از آن باشد. در این صورت $\eta_{\psi\phi} \le \eta_{\psi D} \le \frac{k}{n}$ است.

توجه داشته باشید که در مثال ما، $I_{\psi,\phi}$ تنها تابع توزیع مشترک ψ و ϕ است و ارتباطی با تابع $T = f(\psi)$ ندارد. این بدین معناست که از این نامساوی، می‌توان برای ارائه‌ی کرانی تیز بر روی سوییدگی استفاده کرد.

قضیه پ.۱. اگر $\phi_i - \mu_i$ ها متغیرهای تصادفی σ-زیرگاوسی بوده و $T - \psi - \phi$ برقرار باشد، داریم:

$$\mathbb{E}[\phi_T - \mu_T] \le \sigma\sqrt{2\eta_{\psi\phi} I(T;\psi)} \tag{۱۳۲}$$

اثبات. با توجه به قضیه‌ی (۱.۲) و تعریف نامساوی پردازش اطّلاعات قوی، بدیهی است. $\square$

پ.۲ استفاده‌کردن از اطّلاعات و مسئله‌ی طبقه‌بندی

در این بخش به کاربرد قضیه‌ی مطرح‌شده در این مقاله در کران‌زدن خطای جواب مسئله‌ی طبقه‌بندی در یادگیری ماشین می‌پردازیم. فرض کنید n نمونه‌ی آموزشی در اختیار باشد. هر نمونه، شامل یک بردار $X_1, X_2, \dots, X_n \in \mathcal{X}$ است که هر یک به صورت i.i.d. از توزیع $\mathcal{D}$ انتخاب شده‌اند که توزیع $\mathcal{D}$ در اختیار نمی‌باشد. برای هر یک از این n نمونه، متغیر $Y_i \in \{0,1\}$ داده شده است. فرض کنید که برچسب داده‌ی $X_i = x_i$ از توزیع $\mathbb{P}(Y_i|X_i = x_i)$ گرفته‌شده باشد.

تعریف پ.۲. یک طبقه‌بند، تابعی مانند f از $\mathcal{X}$ به $\{0,1\}$ است. خطای یک طبقه‌بند f بدین صورت تعریف می‌شود:

$$L(f) = \mathbb{E}_{X\sim\mathcal{D}}\left[\mathbf{1}(f(X) \neq Y)\right]$$

خطای طبقه‌بند بر روی داده‌های آموزشی نیز تعریف می‌شود:

$$\hat{L}(f) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(f(x_i) \neq y_i)$$

با این تعریف داریم $\mathbb{E}[\hat{L}(f)] = L(f)$.

الگوریتم یادگیری ماشین به دنبال تابعی مناسب است که رابطه‌ی بین ویژگی‌ها و برچسب‌ها را مدل کند. هدف این است که تابع $\hat{f}$ از یک مجموعه‌ی داده‌شده‌ی توابع مثل $\mathcal{F}$ است به نحوی پیدا شود که $L(\hat{f})$ کمینه‌شود. به دلیل اینکه الگوریتم به توزیع $\mathcal{D}$ دسترسی ندارد، حل این مسئله‌ی بهینه‌سازی به صورت مستقیم ممکن نیست در نتیجه، به کمک روشی دیگر، تابع $\hat{f}$ را انتخاب می‌کنیم. یکی از این روش‌ها، استفاده از

$$\hat{f} = \mathrm{argmin}_{f\in\mathcal{F}}\hat{L}(f)$$

است. به این الگوریتم Empirical Risk Minimization یا ERM می‌گویند.

در حالت کلّی، با چنین روش‌هایی ممکن است دچار مشکلی به نام over-fitting شویم. این به این معنای این است که با وجود کوچک‌بودن $\hat{L}(\hat{f})$ مقدار $L(\hat{f})$ از $\min_{f\in\mathcal{F}} L(f)$ بسیار بزرگ‌تر باشد. تئوری یادگیری ماشین به دنبال بررسی تئوری فاصله‌ی $|L(\hat{f}) - \min_{f\in\mathcal{F}} L(f)|$ است. به طور خاص در تئوری یادگیری ماشین، برای الگوریتم‌ها و مجموعه‌های توابع مختلف، به دنبال تابع $m_{\mathcal{F}}$ در تعریف زیر هستند.

تعریف پ.۳. فرض کنید m نمونه‌ی *i.i.d.* $S = \{(X_1,Y_1),(X_2,Y_2),\dots,(X_m,Y_m)\}$ از توزیع $\mathcal{D}$ در اختیار باشد. توزیع $\mathcal{D}$ دلخواه است. الگوریتم A با دریافت این m نمونه، تابع $\hat{f} = A_S$ از مجموعه‌ی توابع $\mathcal{F}$ را انتخاب می‌کند. می‌گوییم مجموعه‌ی توابع $\mathcal{F}$ توسط الگوریتم A قابل یادگیری به مفهوم PAC هستند، اگر تابع چندجمله‌ای $m_{\mathcal{F}}(.,.)$ وجود داشته‌باشد که به ازای هر $(\epsilon,\delta) \in [0,1]^2$، اگر $m \geq m_{\mathcal{F}}(\epsilon,\delta)$ باشد، با احتمال حداقل $1-\delta$ داشته باشیم:

$$\mathbb{E}_{X\sim\mathcal{D}}\left[\mathbf{1}(A_S(X) \neq Y)\right] \leq \min_{f\in\mathcal{F}}\mathbb{E}_{X\sim\mathcal{D}}\left[\mathbf{1}(f(X) \neq Y)\right] + \epsilon \qquad (۱۳۳)$$

توجه داشته باشید که $m_{\mathcal{F}}(.,.)$ تنها به مجموعه‌ی $\mathcal{F}$ ربط دارد و شرط فوق باید به ازای هر توزیع $\mathcal{D}$ با یک تابع مشخص و ثابت $m_{\mathcal{F}}(.,.)$ برقرار باشد.

مثال پ.۲. مثالی از مسئله‌ی طبقه‌بندی این است که $X_i \in \mathbb{R}^d$ و $\mathcal{F} = \{f_\theta : \theta \in \mathbb{R}^d\}$ باشد که در آن $f_\theta(\mathbf{x}) = \mathbf{1}(\mathbf{x}^\top\theta \geq 0)$ باشد. الگوریتم A نیز θ ای را می‌یابد که در آن $\hat{L}(f_\theta)$ کمینه‌شود. به این مسئله، طبقه‌بندی خطی می‌گوییم. در این مثال، به وضوح می‌دانیم به افزایش d و با ثابت نگاه داشتن تعداد نمونه، ریسک *over-fitting* زیاد می‌شود.

در تئوری یادگیری ماشین، نتایج گوناگونی درباره‌ی کران فاصله‌ی

$$\left|\mathbb{E}_{X\sim\mathcal{D}}\left[\mathbf{1}(A_S(X) \neq Y)\right] - \min_{f\in\mathcal{F}}\mathbb{E}_{X\sim\mathcal{D}}\left[\mathbf{1}(f(X) \neq Y)\right]\right|$$

بر اساس معیارهای پیچیدگی کلاس $\mathcal{F}$ وجود دارد. انتظار داریم با تعداد مشخص نمونه، با افزایش پیچیدگی کلاس $\mathcal{F}$، فاصله‌ی مذکور بزرگ‌تر باشد. یکی از این معیارهای پیچیدگی، بُعد Vapnik-Chervonenkis یا VC Dimension است.

تعریف پ.۴. بُعد VC مجموعه‌ی توابع $\mathcal{F}$ برابر بزرگترین کاردینالیتی مجموعه‌ی $S \subseteq \mathcal{X}$ است به نحوی که توابع $\mathcal{F}$ قادر باشند تمام $2^{|S|}$ برچسب‌گذاری ممکن را بر روی نمونه‌های آن انجام دهند.

تعریف پ.۵. تعداد برچسب‌گذاری‌هایی که توابع مجموعه‌ی $\mathcal{F}$ می‌توانند بر روی یک مجموعه‌ی $S \subseteq \mathcal{X}$ انجام دهند را با $\Pi_{\mathcal{F}}(S)$ نمایش می‌دهیم. تابع رشد مجموعه‌ی توابع $\mathcal{F}$ به این صورت تعریف می‌شود:

$$\Pi_{\mathcal{F}}(m) = \max_{S:\ |S|=m} \Pi_{\mathcal{F}}(S)$$

لم پ.۱. ساور-شلاح: فرض کنید $\mathcal{F}$ مجموعه‌ای از توابع با بعد VC محدود d باشد. در این صورت، به ازای هر $m \in \mathcal{N}$ داریم:

$$\Pi_{\mathcal{F}}(m) \leq (\frac{em}{d})^d \tag{۱۳۴}$$

اثبات. این قضیه در درس تئوری یادگیری ماشین برای ما اثبات شده است و قضیه‌ی (۳.۵) کتاب [۶] است. $\square$

قضیه پ.۲. فرض کنید $\mathbf{x} = (x_1, \dots x_n)$، $\mathbf{Y} = (Y_1, \dots, Y_n)$، $\hat{f}(\mathbf{x}) = (\hat{f}(x_1), \dots \hat{f}(x_n))$ و $\log_+(z) = \max\{1, \log(z)\}$ باشد. در این صورت:

$$\mathbb{E}[L(\hat{f}) - \hat{L}(\hat{f})] \leq \sqrt{\frac{I(\hat{f}(\mathbf{x}); Y)}{2n}} \tag{۱۳۵}$$

به طور خاص، اگر $\mathcal{F}$ دارای بعد VC محدود d باشد، آن‌گاه

$$I(\hat{f}(\mathbf{x}); Y) \leq d \log_+(\frac{ne}{d}) \tag{۱۳۶}$$

نتیجه پ.۱. در تعریف یادگیری PAC دیدیم که برای این نوع یادگیری، نیاز است کرانی ثابت برای تمام توزیع‌های ممکن ورودی به دست آوریم. کران‌های معمول در تئوری یادگیری ماشین نیز از این جنس هستند که مستقل از توزیع ورودی، برای هر توزیع دلخواهی خطا را کران می‌زنند. از طرفی، کران (۱۳۵) وابسته به توزیع داده‌هاست. این نوع کران‌ها می‌توانند منجر به فهم بهتر این مسئله شوند که چه نوع توزیع‌هایی به نحوی بهتر قابل یادگیری هستند. همچنین برای برخی توزیع‌ها،کران به دست آمده، می‌تواند بسیار تیز از کران‌های نظریه‌ی PAC باشد.

پ.۳ مشاهده‌ی داده‌ها و تعداد کلاس‌ها در خوشه‌بندی

در بسیاری از مواقع، در مسائل خوشه‌بندی داده‌ها با روش‌هایی مثل K-Means، شخص پردازش‌گر داده، با مشاهده‌ی نمودارهای داده، تعداد خوشه‌ها، K، را مشخص می‌کند. در این تحلیل و با توجه به قضیه‌ی (۱.۲)، سوییدگی ناشی از این کار، اگر $I(T; \phi)$ کوچک باشد، کم است. بر اساس نامساوی پردازش اطّلاعات داریم: $I(T; \phi) \leq I(K; \phi)$ است. در مواردی که ساختار مشخضی در داده وجود دارد که الگوریتم خوشه‌بندی قابل درک آن است، معمولاً Kحول مقادیر خاصی متمرکز است و $I(K; \phi) \leq H(K) \approx 0$ است. در این مثال، تعداد خوشه‌ها برای تحلیل‌گر داده مفید است ولی مصداق «استفاده‌ی بد از اطّلاعات» نیست.

پ.۴ کنترل سوییدگی از طریق کنترل FDR

کنترل False Discovery Rate یکی از مسائل مطرح در هنگام انجام تعداد زیاری آزمون‌های فرضیه همزمان است. فرض کنید دیتاست بزرگ $D \in \mathbb{R}^{n \times m}$ شامل n نمونه از m ژن باشد. فرض کنید n_1 نمونه‌ی اول از بافتی سرطانی بوده و $n - n_1$ نمونه‌ی بعدی نیز از بافت سالم آمده باشند. اصولاً دانشمندان علاقه‌مند به یافتن ژن‌هایی هستن که در توده‌ی سرطانی و توده‌های سالم تفاوت معناداری داشته‌باشند. برای این کار، برای هر ژن، یک آزمون فرضیه انجام می‌شود که رد شدن فرضیه‌ی صفر آن، نشان دهنده‌ی این است که بین توزیع این ژن در بافت سرطانی و بافت سالم تفاوت وجود دارد. در این حالت، بسیار نامحتمل است

تفاوت بین ژن در بافت‌های سرطانی و سالم به صورت تصادفی بوده باشد. هدف تنظیم آستانه‌ی رد و قبول فرضیه در این آزمون فرض است.

در این‌جا به بررسی یک حالت بسیار کلّی از این مسئله می‌پردازیم. فرض کنید $D \in \mathbb{R}^{n\times m}$ یک ماتریس تصادفی بوده و بردار $\phi \in \mathbb{R}^m$ تابعی از آن باشد. به عنوان مثال، این تابع می‌تواند آماره‌های ستون‌های ماتریس D باشد، به عنوان مثال، تفاوت میانگین ستون i ام در بافت‌های سرطانی و بافت‌های سالم. امید ریاضی این بردار را μ بگیرید، یعنی $\mu = \mathbb{E}\phi(D)$. اندیس‌های $\{1,\ldots,m\}$ به دو بخش H_0 و H_1 تقسیم شده‌اند. یک فرآیند انتخاب، تابعی مثل $\psi : \mathbb{R}^{n\times m} \to \{0,1\}^m$ است که $\psi(D)_i = 1$ نشان می‌دهد که ویژگی (یا ژن) i ام انتخاب شده است. مجموعه‌ی $S_1 \subseteq \{1,\ldots,m\}$ را مجموعه‌ی ژن‌های انتخاب شده و مجموعه‌ی S_2 را مجموعه‌ی ژن‌های انتخاب نشده در نظر بگیرید. در این فرمول‌بندی، ستون‌های با اندیس عضو H_0، ستون‌هایی هستند که فرض صفر برای آن‌ها برقرار است و تفاوت ژن مربوطه در بافت سالم و سرطانی، معنادار نیست. تعریف‌های زیر، تعریف‌هایی مشابه تعریف های معمول خطای نوع اول و خطای نوع دوم هستند:

$$\hat\alpha = \frac{\#(H_0 \cap S_1)}{\#H_0} \qquad \hat\beta = \frac{\#(H_1 \cap S_0)}{\#H_1}$$

ما علاقه‌مند هستیم که متغیّرهایی به فرم زیر را بررسی کنیم:

$$\frac{1}{\#S_1}\sum_{i\in S_1}(\phi_i - \mu_i) \tag{۱۳۷}$$

$$\frac{1}{\#S_1}\sum_{i\in S_1}|\phi_i - \mu_i| \tag{۱۳۸}$$

$$\frac{1}{\#S_1}\sum_{i\in S_1}(\phi_i - \mu_i)^2 \tag{۱۳۹}$$

این متغیرها، به معنای خطا (یا سوییدگی) میانگین در متغیرهای انتخاب‌شده هستند. این کمیت‌ها را می‌توان به ترتیب به صورت $\mathbb{E}[\phi_T - \mu_T]$، $\mathbb{E}[\phi_T - \mu_T]$ و $\mathbb{E}[\phi_T - \mu_T]$ نوشته‌شوند که در آن، متغیر تصادفی T به شرط D دارای توزیع یکنواخت بر روی متغیرهای انتخاب شده در S_1 دارد.

حال، FDR را به صورت $\mathrm{FDR} = \mathbf{P}(T \in H_0)$ تعریف می‌کنیم. این متغیر، نشان می‌دهد که برای چه کسری از متغیرهای انتخاب شده، عضو H_0 بوده‌اند و اشتباه رخ داده است. قضیه‌ی فوق قضیه‌ای بسیار مهم است که نشان می‌دهد که در صورتی که در فاز آزمون فرضیه و انتخاب ویژگی خوب عمل کنیم و FDR کوچکی داشته باشیم، اطّلاعات متقابل $I(T;\phi)$ نیز کوچک است و در نتیجه در فاز تخمین نیز دچار سوییدگی بزرگی نخواهیم بود.

قضیه پ.۳. در مسئله‌ی فوق، داریم:

$$I(T;\phi) \le h(\mathrm{FDR}) + (1-\mathrm{FDR})\log(\frac{1}{1-\beta}) + \mathrm{FDR}\log(\frac{1}{\alpha}) + \xi \tag{۱۴۰}$$

که در آن، $h(p)$ تابع آنتروپی باینری است. $\alpha = \mathbb{E}\hat\alpha$ و $\beta = \mathbb{E}\hat\beta$ خطاهای توع اول و دوم هستند و داریم:

$$\xi = \mathbb{E}\Big[\log_+(\frac{1-\beta}{1-\hat\beta})\Big] + \mathbb{E}\Big[\log_+(\frac{\alpha}{\hat\alpha})\Big]$$

اثبات. تعریف می‌کنیم $\mathcal{X} = \mathbf{1}(T \in H_0)$. از آن‌جایی که $\mathcal{X}$ تابعی یقینی از T است، می‌توان نوشت:

$$I(T;\phi) = I(T;\mathcal{X},\phi) = I(T;\mathcal{X}) + I(T;\phi|\mathcal{X}) \le H(\mathcal{X}) + I(T;\phi|\mathcal{X}) \tag{۱۴۱}$$

با استفاده از توزیع $\mathcal{X}$ و استفاده از نامساوی پردازش اطّلاعات و قاعده‌ی زنجیر، داریم:

$$I(T;\phi|\mathcal{X}) \le H(\mathcal{X}) + \mathbf{P}(\mathcal{X}=0)I(T;\phi|\mathcal{X}=0) + \mathbf{P}(\mathcal{X}=1)I(T;\phi|\mathcal{X}=1) \tag{۱۴۲}$$

$$= h\big(\mathbf{P}(T\in H_1)\big) + \mathbf{P}(T\in H_1)I(T;\phi|T\in H_1) \tag{۱۴۳}$$

$$+ \mathbf{P}(T\in H_0)I(T;\phi|T\in H_0) \tag{۱۴۴}$$

$$\le h(\mathbf{P}(T\in H_1)) + \mathbf{P}(T\in H_1)I(T;D|T\in H_1) \tag{۱۴۵}$$

$$+ \mathbf{P}(T\in H_0)I(T;D|T\in H_0) \tag{۱۴۶}$$

می‌توان به سادگی $I(T; D \mid T \in \mathcal{H}_1)$ را به فرم زیر کران زد:

$$
\begin{aligned}
I(T; D \mid T \in \mathcal{H}_1) &= H(T \mid T \in \mathcal{H}_1) - H(T|T \in \mathcal{H}_1,\, D) && (۱۴۷)\\
&\leq \log(\#\mathcal{H}_1) - \mathbb{E}[\log\left(\#(S_1 \cap \mathcal{H}_1)\right) \mid T \in \mathcal{H}_1] && (۱۴۸)\\
&= \log(\#\mathcal{H}_1) - \mathbb{E}[\log\left((1-\hat{\beta}) \cdot (\#\mathcal{H}_1)\right) \mid T \in \mathcal{H}_1] && (۱۴۹)\\
&= -\mathbb{E}\left[\log\left(1-\hat{\beta}\right) \mid T \in \mathcal{H}_1\right] && (۱۵۰)\\
&= -\log(1-\beta) + \mathbb{E}\left[\log\left(\frac{1-\beta}{1-\hat{\beta}}\right) \mid T \in \mathcal{H}_1\right] && (۱۵۱)
\end{aligned}
$$

با استفاده از این کران، داریم:

$$
\begin{aligned}
\mathbb{P}(T \in \mathcal{H}_1) I(T; D \mid T \in \mathcal{H}_1) \leq{}& -\mathbb{P}(T \in \mathcal{H}_1)\log(1-\beta) && (۱۵۲)\\
&+ \mathbb{E}\left[\log\left(\frac{1-\beta}{1-\hat{\beta}}\right) \mathbf{1}_{\{T \in \mathcal{H}_1\}}\right] && (۱۵۳)\\
\leq{}& -\mathbb{P}(T \in \mathcal{H}_1)\log(1-\beta) && (۱۵۴)\\
&+ \mathbb{E}\left[\log_+\left(\frac{1-\beta}{1-\hat{\beta}}\right)\right] && (۱۵۵)
\end{aligned}
$$

با تکرار محاسبات فوق، به نتیجه‌ی زیر می‌رسیم:

$$
\mathbb{P}(T \in \mathcal{H}_0) I(T; X \mid T \in \mathcal{H}_0) \leq -\mathbb{P}(T \in \mathcal{H}_0)\log(\alpha) + \mathbb{E}\left[\log_+\left(\frac{\alpha}{\hat{\alpha}}\right)\right] \qquad (۱۵۶)
$$

با جای‌گذاری این نتایج در معادله‌ی (۱۴۱)، قضیه اثبات می‌شود. $\square$