دانشگاه رازی

دانشکده علوم
گروه آمار

پایان نامه کارشناسی ارشد آمار ریاضی

کران‌های بهبود یافته برای معیار کولبک — لیبلر بر اساس ترکیب مدل‌های رقابتی

توسط

طیبه کریمی

استاد راهنما

دکترعبدالرضا سیاره

استاد مشاور

تیر ماه ۱۳۹۰

کران‌های بهبود یافته برای معیار کولبک – لیبلر بر اساس ترکیب مدل‌های رقابتی

چکیده

یکی از مفاهیم بنیادی در استنباط آماری انتخاب مدل مناسب برای یک مجموعه از داده‌ها است. هنگامی که مجموعه‌ای از داده‌ها در اختیار ما قرار می‌گیرند چگالی مولد این داده‌ها یعنی تابع چگالی درست داده‌ها مجهول است. لذا با مجموعه‌ای از مدل‌های رقابتی روبرو خواهیم بود. انتخاب یک مدل قطعی از بین این مدل‌های رقابتی که براساس تعداد محدودی از مشاهدات پیشنهاد شده‌اند، به عنوان برآوردی از چگالی درست جامعه موجب بروز ریسک در انتخاب مدل برای جامعه خواهد شد. به همین دلیل انتخاب مدل بهینه از بین این مدل‌های رقابتی هدف اصلی انتخاب مدل است. معیارها و آزمون‌های مختلفی برای انتخاب مدل بهینه معرفی شده‌اند. معیار واگرایی کولبک – لیبلر با کاربرد گسترده در ساختار این معیارها و آزمون‌ها مورد توجه ما است. چون معیار کولبک – لیبلر واگرا است، پیدا کردن کران‌های مناسب بالایی و پایینی برای این معیار، کمک خواهد کرد تا مدل مناسبی به عنوان برآورد مدل درست انتخاب شود. در این پایان نامه به بررسی کران‌های موجود برای معیار اطلاع کولبک – لیبلر پرداخته شده است. سپس خواص شکل نمایی این معیار مورد بررسی قرار گرفته و به کمک این خواص و ترکیب محدب مدل‌های رقابتی، این معیار بهبود داده شده است. همچنین به بررسی و مقایسه معیار اطلاع کولبک – لیبلر میانگین‌های حسابی، هارمونیک و هندسی مدل‌های رقابتی پرداخته شده است.

واژه‌های کلیدی : انتخاب مدل، ترکیب محدب، خاصیت زبر جمعی، معیار اطلاع، معیار کولبک – لیبلر، میانگین هندسی وزنی.

فهرست مندرجات

فصل ۱

تعاریف و مفاهیم اولیه

۱—۱ مقدمه

مدل‌های آماری نقش بسیار مهمی در تحلیل و آنالیز داده‌ها دارند. در دنیای واقعی با رویدادهای پیچیده‌ای روبرو هستیم که برای تحلیل روند آنها و پیش بینی رفتار آینده باید بتوانیم الگوی آنها را بیابیم. در علم آمار این الگوها را با مدل‌های آماری نشان می‌دهند. مدل‌های آماری برای نشان دادن ساختارهای تصادفی، پیش بینی رفتار آینده و استخراج اطلاعات مهم از داده‌ها مورد استفاده قرار می‌گیرند.

آکائیک[۱] (۱۹۸۵ و ۱۹۷۴) این نقطه نظر را مطرح کرد که مدل سازی آماری فقط پیدا کردن مدلی نیست که رفتار داده‌های مشاهده شده را شرح دهد، بلکه هدف اصلی آن پیش بینی تا حد امکان خوب آینده فرایند تحت بررسی است. یک نقطه نظر دیگر در مدل‌های آماری استخراج اطلاعات از داده‌ها است. در بسیاری از استنباط‌های آماری فرض را بر این می‌گذارند که داده‌ها از یک مدل مشخص پیروی می‌کنند. بر همین اساس یک مدل آماری پارامتری را پیشنهاد کرده و به دنبال برآورد مدل درست مشاهدات هستند. اما در دنیای واقعی مشاهدات و رخدادها تحت تاثیر متغیرهای زیادی هستند که بعضی از آنها قابل شناسایی نیستند. پس پیدا کردن مدلی که به طور کاملا دقیق توزیع مشاهدات را نشان دهد کار بسیار پیچیده و گاهی غیر ممکن است. لذا سعی می‌شود مدلی مورد استفاده قرار گیرد که در میان مدل‌های رقابتی که برای برآورد مدل درست پیشنهاد شده اند، مدل بهینه باشد.

یک مدل خوب به صورت یکتا قابل تعیین کردن نیست و می‌تواند بسته به نظر آماردان یا اطلاعات موجود متفاوت باشد. به عبارت دیگر هدف مدل سازی آماری ساختن یک مدل یکتا نیست بلکه ساختن یک مدل خوب تحت شرایط موجود است.

معیارهای متفاوتی برای انتخاب مدل معرفی شده‌اند، یکی از معیارهایی که کاربرد فراوانی در این زمینه دارد معیار اطلاع کولبک — لیبلر[۲] است. اطلاع کولبک — لیبلر در سال ۱۹۵۱ توسط سالمون کولبک[۳] و ریچارد لیبلر[۴] برای سنجش میزان نزدیکی مدل انتخابی به مدل درست معرفی شد.

از آنجا که این معیار کاربرد گسترده‌ای در انتخاب مدل دارد، انتخاب حدود بالایی و پایینی مناسب برای این معیار در انتخاب یک مدل مناسب برای داده‌ها به ما کمک می‌کند.

در این پایان نامه به بررسی کران‌هایی برای اطلاع کولبک — لیبلر بخصوص کران‌های بالایی برای آن پرداخته شده و سپس برخی خواص نمایی این معیار بررسی شده است. همچنین این معیار را در حالتی

[۱] Akaike

[۲] Kullback-Leibler Information

[۳] Solomon Kullback

[۴] Richard Leibler

که مدل‌های رقابتی ترکیب شده‌اند مورد بررسی قرار داده و مقایسه‌ای بین ترکیب های مختلف به عمل آمده است.

در فصل اول تعاریف، قضایا و نامساوی‌های مورد نیاز فصل‌های بعد و معیار اطلاع کولبک — لیبلر معرفی می‌شوند. در فصل دوم کران‌های پایینی و بالایی برای اطلاع کولبک — لیبلر آورده شده است. در فصل سوم خواص شکل نمایی اطلاع کولبک — لیبلر بیان می‌شود و در فصل چهارم ترکیب محدب مدل‌های رقابتی را به عنوان مدل پیشنهادی معرفی کرده و اطلاع کولبک — لیبلر در این حالت بررسی می‌شود و مقایسه‌ای بین ترکیب‌های مختلف مدل‌های رقابتی به عمل خواهد آمد.

۱—۲ تعاریف وقضایا

در این بخش برخی تعاریف، قضایا و نامساوی‌های مورد نیاز فصل‌های بعدی بیان شده است.

تعریف ۱.۱ میانگین هارمونیک[۵]

فرض کنید $a = (a_۱, a_۲, ..., a_n)$ برداری از اعداد حقیقی مثبت باشد، آنگاه میانگین هارمونیک به صورت

$$H(a_۱, a_۲, ..., a_n) = \frac{n}{\sum_{i=۱}^{n} \frac{۱}{a_i}}$$

تعریف می‌شود.

تعریف ۲.۱ میانگین هندسی وزنی[۶]

فرض کنید $p = (p_۱, p_۲, ..., p_n)$ توزیع احتمال و $a = (a_۱, a_۲, ..., a_n)$ برداری از اعداد حقیقی نامنفی باشند، میانگین هندسی وزنی که با نماد $G_n(p, a)$ نشان داده می‌شود، به صورت

$$G_n(p, a) = \prod_{i=۱}^{n} a_i^{p_i}$$

تعریف می‌شود.

---

۵ Harmonic Mean

۶ Weighted Geometric Mean

تعریف ۳.۱ تابع محدب[۷]

تعاریف مختلفی از توابع محدب وجود دارد که به سه نوع از این تعاریف اشاره می‌کنیم:

الف) تابع پیوسته $g(.)$ که دامنه و برد آن اعداد حقیقی است، محدب نامیده می‌شود اگر برای هر $x_\circ \in \Re$، خطی مانند $l(x)$ وجود داشته باشد که از نقطه $(x_\circ, g(x_\circ))$ بگذرد و رابطه

$$l(x) \leq g(x), \quad \forall x \in \Re \tag{۱.۲.۱}$$

برقرار باشد.

ب) برای مقادیر حقیقی تابع اکیداً محدب $g(.)$ که روی فاصله حقیقی I تعریف می‌شود نامساوی

$$g(b) - g(a) \geq g'(a)(b-a) \tag{۲.۲.۱}$$

برای هر $a, b \in I$ برقرار است. هرگاه $a = b$ باشد در نامساوی فوق تساوی رخ می‌دهد.

ج) تابع پیوسته $g(.)$ اکیداً محدب است اگر مشتق دوم آن همواره مثبت باشد.

قضیه ۱.۱ . نامساوی جنسن[۸]

اگر X متغیری تصادفی با میانگین $E[X]$ و $g(.)$ تابعی محدب باشد، آنگاه

$$E[g(X)] \geq g(E[X]). \tag{۳.۲.۱}$$

برهان. اگردر تعریف الف (۱.۳) فرض کنیم $x_\circ = E[X]$ با توجه به تحدب و پیوستگی $g(.)$ خطی مانند $l(x) = a + bx$ وجود دارد به طوری که

$$l(E[X]) = g(E[X]) \tag{۴.۲.۱}$$

لذا طبق رابطه (۱.۲.۱)

---

۷ Convex Function

۸ Jensen Inequality

$$l(x) = a + bx \leq g(x), \qquad \forall x \in \Re \tag{۵.۲.۱}$$

از طرفی

$$E[l(X)] = E[a + bX] = a + bE[X] = l(E[X]) \tag{۶.۲.۱}$$

لذا طبق رابطه فوق و رابطه‌های (۴.۲.۱) و (۵.۲.۱) نامساوی

$$g(E[X]) = l(E[X]) = E[l(X)] \leq E[g(X)] \tag{۷.۲.۱}$$

نتیجه می‌شود.

تذکر: برای اثبات نامساوی (۷.۲.۱) از این ویژگی امید ریاضی، که اگر $f(x) \leq h(x)$ آنگاه $E[f(X)] \leq E[h(X)]$ استفاده شده است. □

۱—۳ مدل‌های آماری

فرض می کنیم $\{X_1, X_2, ..., X_n\}$ یک نمونه تصادفی $i.i.d$ از توزیعی با تابع چگالی $h(x)$ باشد. $h(x)$ تابع چگالی درست یا مدل درست مشاهدات است. این توزیع مجهول است و برآورد آن یکی از مسائل اساسی در آمار است. با توجه به ساختار داده‌ها، مدلی مانند $F(x)$ برای تقریب یا برآورد توزیع درست معرفی می‌شود. تابع چگالی احتمال مرتبط با این مدل با $f(x)$ نشان داده می‌شود. اگر هدف برآورد مدل درست باشد معمولا مدل پارامتری $F_\theta = \{f(x;\theta), \theta \in \Theta \subset \Re^p\}$ به عنوان یک مدل رقابتی انتخاب خواهد شد. با برآورد پارامترهای مدل رقابتی، برآوردی برای مدل درست به دست می آید.

مدل‌های آماری نسبت به هم می‌توانند آشیانه‌ای [۹] ، غیر آشیانه‌ای[۱۰] و متداخل[۱۱] باشند.

تعریف ۴.۱ مدل‌های غیر آشیانه‌ای .

فرض کنید دو مدل $F_\theta = \{f(x;\theta), \theta \in \Theta \subset \Re^p\}$ و $G_\gamma = \{g(x;\gamma), \gamma \in \Gamma \subset \Re^q\}$ مدل‌های رقابتی

[۹] Nested

[۱۰] Non - nested

[۱۱] Overlap

برای چگالی درست باشند، این دو مدل را نسبت به هم غیر آشیانه‌ای گوییم اگر و تنها اگر $F_\theta \bigcap G_\gamma = \emptyset$.

تعریف ۵.۱ مدل‌های آشیانه‌ای .

گوییم مدل F_θ در مدل G_γ آشیانه دارد هرگاه $F_\theta \subset G_\gamma$.

تعریف ۶.۱ مدل‌های متداخل .

دو مدل F_θ و G_γ را متداخل گوییم اگر و تنها اگر:

۱) $F_\theta \bigcap G_\gamma \neq \emptyset$

۲) $G_\gamma \nsubseteq F_\theta$ و $F_\theta \nsubseteq G_\gamma$

مدل‌های آماری را می‌توان در دو دسته طبقه‌بندی کرد: مدل‌های خوب — توصیف شده[۱۲] و مدل‌های بد — توصیف شده[۱۳] .

تعریف ۷.۱ مدل خوب — توصیف شده.

مدل $F_\theta = \{f(x;\theta), \theta \in \Theta \subset \Re^p\}$ را خوب — توصیف شده گوییم هرگاه یک $\theta_\circ \in \Theta$ وجود داشته باشد که $h \in F_\theta$. به عبارت دیگر $h(x) = f(x;\theta_\circ)$.

تعریف ۸.۱ مدل بد — توصیف شده.

مدل F_θ را بد — توصیف شده گوییم هرگاه $\theta_\circ \in \Theta$ وجود نداشته باشد که $h \in F_\theta$. به عبارت دیگر برای هر $\theta \in \Theta$، $h(x) \neq f(x;\theta)$.

۱—۴ معیار اطلاع کولبک — لیبلر

معیار اطلاع کولبک — لیبلر یک معیار برای ارزیابی میزان نزدیکی مدل‌های آماری به توزیع درست جامعه است. معیار اطلاع KL در واقع میزان ریسک احتمالی انتخاب یک مدل رقابتی به جای توزیع درست است. هرچه این معیار کوچکتر باشد مدل انتخابی مدلی مناسب‌تر خواهد بود.

فرض می‌کنیم که $f(x)$ یک مدل آماری ساخته شده بر اساس مشاهدات باشد. می‌خواهیم نزدیکی این مدل را به تابع چگالی (تابع جرم احتمال) درست $h(x)$ بر اساس معیار اطلاع KL ارزیابی کنیم. این معیار به صورت

---

۱۲ Well-Specified

۱۳ Miss-Specified

$$KL(h,f) = E_h\Big[\log\{\tfrac{h(X)}{f(X)}\}\Big]$$

است که در آن E_h امید ریاضی تحت تابع چگالی یا تابع جرم احتمال h است.

اگر h یک تابع چگالی پیوسته باشد این معیار به صورت

$$KL(h,f) = \int_{-\infty}^{+\infty} \log\{\frac{h(x)}{f(x)}\} h(x)dx \tag{۱.۴.۸}$$

تعریف می‌شود. اگر h یک تابع احتمال باشد آنگاه

$$KL(h,f) = \sum_{i=۱}^{n} f(x_i) \log\{\frac{h(x_i)}{f(x_i)}\}. \tag{۱.۴.۹}$$

۱—۴—۱ ویژگی‌های معیار اطلاع کولبک — لیبلر

معیار اطلاع کولبک — لیبلر به عنوان معیار واگرایی بین دو توزیع احتمال معرفی شد. این معیار یک متر بر روی فضای توزیع‌های احتمال نیست زیرا:

۱) اطلاع کولبک — لیبلر متقارن نیست یعنی

$$KL(h,f) \neq KL(f,h)$$

۲) این معیار در نامساوی مثلثی [۱۴] صدق نمی‌کند.

$$KL(h,f) \nleq KL(h,g) + KL(g,f)$$

اطلاع کولبک — لیبلر دارای ویژگی های زیر است:

۱) $KL(h,f) \geq ۰$

۲) $KL(h,f) = ۰$ اگر و تنها اگر $h(x) = f(x)$.

برهان.

برای بررسی دو ویژگی بالا، تابع $k(t) = \log t - t + ۱$ را که برای $t > ۰$ تعریف شده است در نظر می‌گیریم. در این صورت مشتق $k(t)$ برابر است با

$$k'(t) = \frac{۱}{t} - ۱,$$

[۱۴] Triangle Inequality

که $k'(۱) = ۰$ و همچنین

$$k''(t) = -\frac{۱}{t^۲} < ۰,$$

در نتیجه $k(t)$ ماکسیمم مقدار خود را در نقطه یک اختیار می کند، $k(۱) = ۰$.

بنابراین

$$k(t) \leq k(۱) = ۰, \quad \forall t > ۰$$

لذا

$$\log(t) \leq t - ۱, \quad \forall t > ۰$$

و تساوی زمانی برقرار است اگر و فقط اگر $t = ۱$ باشد.

اثبات ویژگی ۱:

الف) برای حالت پیوسته با جایگذاری $t = \frac{f(x)}{h(x)}$ داریم:

$$\log(\frac{f(x)}{h(x)}) \leq (\frac{f(x)}{h(x)}) - ۱,$$

با ضرب $h(x)$ در دو طرف نامساوی فوق و انتگرال گرفتن خواهیم داشت:

$$\begin{aligned} \int_{\Re} \log\{\frac{f(x)}{h(x)}\} h(x) dx &\leq \int_{\Re} \{\frac{f(x)}{h(x)} - ۱\} h(x) dx \\ &= \int_{\Re} \{f(x) - h(x)\} dx \\ &= \int_{\Re} f(x) dx - \int_{\Re} h(x) dx \end{aligned} \qquad (۱۰.۴.۱)$$

چون انتگرال هر تابع چگالی روی فضای متغیر تصادفی X برابر یک است در نتیجه طرف دوم رابطه فوق صفر خواهد شد، بنابراین:

$$\int_{\Re} \log\{\frac{f(x)}{h(x)}\} h(x) dx \leq ۰,$$

و

$$\begin{aligned} KL(h, f) &= \int_{\Re} \log\{\frac{h(x)}{f(x)}\} h(x) dx \\ &= -\int_{\Re} \log\{\frac{f(x)}{h(x)}\} h(x) dx \geq ۰ \end{aligned} \qquad (۱۱.۴.۱)$$

ب) در حالت گسسته تابع جرم احتمال به جای تابع چگالی و سیگما به جای انتگرال را می‌توان در نظر گرفت.

اثبات ویژگی ۲:

همانطور که گفته شد در نامساوی $\log(t) \le t - ۱$ تساوی برقرار است اگر و تنها اگر $t = ۱$ باشد. با توجه با اینکه در اثبات ویژگی ۱، $t = \frac{f(x)}{h(x)}$ را فرض کردیم ویژگی ۲ برقرار است.

براساس ویژگی‌های فوق در می‌یابیم که هر چه معیارکولبک — لیبلر کوچک‌تر باشد مدل مورد نظر به چگالی درست مشاهدات نزدیک‌تر است و تنها در صورتی برابر صفر خواهد شد که مدل رقابتی معادل چگالی درست مشاهدات باشد.

در واقع $\log\{\frac{h(X)}{f(X)}\}$ تابع زیان حاصل از انتخاب مدل $f(x)$ به جای مدل درست $h(x)$ است و امید ریاضی آن نسبت به چگالی درست $h(x)$ ریسک کولبک — لیبلر خواهد بود. این ریسک وقتی به صفر نزدیک‌تر می‌شود که مدل ساخته شده تقریبا معادل چگالی درست مشاهدات باشد. لذا اطلاع کولبک — لیبلر، ریسک کولبک — لیبلر نیز نامیده می‌شود.

در مثال‌های زیر معیار کولبک — لیبلر برای چند مدل پیوسته و گسسته بدست آمده است.

معیار کولبک — لیبلر برای حالت پیوسته

مثال ۱.۱ فرض کنید مدل درست h و مدل رقابتی f به ترتیب دارای توزیع $N(\zeta, \tau^۲)$ و $N(\mu, \sigma^۲)$ در نظر گرفته شده باشند. برای توزیع $N(\mu, \sigma^۲)$:

$$f(x) = (۲\pi\sigma^۲)^{(-۱/۲)} \exp\{-\frac{(x-\mu)^۲}{۲\sigma^۲}\}$$

لذا

$$\begin{aligned} E_h[\log f(X)] &= E_h[-\frac{۱}{۲}\log(۲\pi\sigma^۲) - \frac{(X-\mu)^۲}{۲\sigma^۲}], \\ &= -\frac{۱}{۲}\log(۲\pi\sigma^۲) - \frac{\tau^۲ + (\zeta-\mu)^۲}{۲\sigma^۲}. \qquad (۱۲.۴.۱) \end{aligned}$$

و

$$E_h[\log h(X)] = -\frac{۱}{۲}\log(۲\pi\tau^۲) - \frac{۱}{۲}. \qquad (۱۳.۴.۱)$$

در نتیجه اطلاع کولبک — لیبلر به صورت

$$
\begin{aligned}
KL(h,f) &= E_h[\log h(X)] - E_h[\log f(X)] \\
&= \frac{۱}{۲}\Big\{ \log(\frac{\sigma^۲}{\tau^۲}) + \frac{\tau^۲ + (\zeta-\mu)^۲}{\sigma^۲} - ۱ \Big\} \qquad (۱۴.۴.۱)
\end{aligned}
$$

خواهد بود.

مثال ۲.۱ فرض کنید که مدل درست جامعه توزیع لاپلاس با تابع چگالی $h(x) = \frac{۱}{۲}\exp(-|x|)$ و مدل رقابتی $f(x)$ نرمال با میانگین μ و واریانس $\sigma^۲$ باشد. در این حالت داریم:

$$
\begin{aligned}
E_h[\log h(X)] &= -\log ۲ - \frac{۱}{۲}\int_{-\infty}^{+\infty} |x| e^{-|x|} dx \\
&= -\log ۲ - \int_{۰}^{+\infty} x e^{-x} dx \\
&= -\log ۲ - ۱ \qquad (۱۵.۴.۱)
\end{aligned}
$$

$$
\begin{aligned}
E_h[\log f(X)] &= -\frac{۱}{۲}\log(۲\pi\sigma^۲) - \frac{۱}{۴\sigma^۲}\int_{-\infty}^{+\infty} (x-\mu)^۲ e^{-|x|} dx \\
&= -\frac{۱}{۲}\log(۲\pi\sigma^۲) - \frac{۱}{۴\sigma^۲}(۴ + ۲\mu^۲) \qquad (۱٦.۴.۱)
\end{aligned}
$$

اطلاع کولبک – لیبلر برای مدل $f(x)$ تحت چگالی درست $h(x)$ برابراست با:

$$
\begin{aligned}
KL(h,f) &= E_h[\log h(X)] - E_h[\log f(X)] \\
&= \frac{۱}{۲}\log(۲\pi\sigma^۲) + \frac{۲+\mu^۲}{۲\sigma^۲} - \log ۲ - ۱ \qquad (۱۷.۴.۱)
\end{aligned}
$$

مثال ۳.۱ معیارکولبک – لیبلر برای حالت گسسته

در پرتاب یک تاس سالم، احتمال آمدن هریک از خال‌ها برابر $\frac{۱}{٦}$ است. لذا تابع جرم احتمال درست به

صورت $p=\{\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6}\}$ است. دو تابع جرم احتمال برای مشاهده خال‌های این تاس به شکل زیر پیشنهاد شده است:

$$q_1=\{0/20, 0/12, 0/18, 0/12, 0/20, 0/18\}$$

و

$$q_2=\{0/18, 0/12, 0/14, 0/19, 0/22, 0/15\}$$

می‌خواهیم بررسی کنیم کدام مدل برای توصیف نتایج حاصل از پرتاب یک تاس سالم مناسب‌تر است. بااستفاده از تعریف اطلاع کولبک – لیبلر برای حالت گسسته داریم:

$$KL(p,q_1)=\sum_{i=1}^{6}p_i\log\{\frac{p_i}{q_{1i}}\}=0/023$$

و

$$KL(p,q_2)=\sum_{i=1}^{6}p_i\log\{\frac{p_i}{q_{2i}}\}=0/020$$

نامنفی بودن اطلاع کولبک – لیبلر در این مثال به وضوح دیده می‌شود. همچنین با توجه به مقادیر بدست آمده، احتمال‌های متناظر با مدل q_2 به مدل درست نزدیک‌تر است. بنابراین مدل q_2 به مدل q_1 ترجیح داده می‌شود.

۱–۵ معیار اطلاع کولبک – لیبلر پارامتری

معیار اطلاع کولبک – لیبلر در حالت پارامتری به صورت

$$KL(h,f_\theta)=E_h\Big[\log\Big(\frac{h(X)}{f(X;\theta)}\Big)\Big]$$

است. این معیار می‌تواند به صورت زیر تجزیه شود:

$$KL(h,f_\theta)=E_h\big[\log h(X)\big]-E_h\big[\log f(X;\theta)\big]$$

به وضوح دیده می‌شود که معیار اطلاع کولبک – لیبلر تابعی از پارامتر θ است. چون جمله اول سمت راست رابطه فوق یک مقدار ثابت است و فقط به چگالی درست و مجهول h بستگی دارد، بنابراین برای مقایسه مدل‌های رقابتی کافیست عبارت دوم سمت راست را که جمله مرتبط معیار اطلاع کولبک – لیبلر نامیده می‌شود، محاسبه کرد. از آنجایی که در بحث انتخاب مدل علاقمند به پیدا کردن θ ایی

هستیم که معیار KL را مینیمم کند، کافیست θ ایی را پیدا کنیم که جمله مرتبط را ماکسیمم کند. توجه کنید جمله مرتبط در واقع امید لگاریتم درستنمایی است. بنابراین در مقایسه مدل‌های رقابتی مدلی را بر سایر مدل‌ها ترجیح می‌دهیم که امید لگاریتم درستنمایی بزرگتری داشته باشد.

تعریف ۹.۱ مقدار شبه پارامتر درست [۱۵]

فرض کنید $F_\theta = \{f(x;\theta), \theta \in \Theta \subset \Re^p\}$ مدل پیشنهاد شده باشد. پارامتر θ_* ایی که معیار کولبک – لیبلر را مینیمم کند، مقدار شبه پارامتر درست نامیده می‌شود.

با توجه به مجهول بودن چگالی درست جامعه h و پارامتر θ، مقدار امید لگاریتم درستنمایی یک کمیت مجهول است بنابراین باید برآورد شود تا بتوان از آن برای انتخاب مدل استفاده کرد.

۱–۶ معیار اطلاع آکائیک (AIC)

فرض کنید $\{X_۱, X_۲, ..., X_n\}$ یک نمونه تصادفی از چگالی درست و مجهول h و یک خانواده F_θ از مدل‌های رقابتی برای برآورد h در نظر گرفته شده باشند. با استفاده از روش درستنمایی ماکسیمم، برآورد پارامترها را پیدا کرده و چگالی $f(x, \hat{\theta}_n)$ تشکیل داده می‌شود. در اینجا هدف محاسبه و بررسی خوبی یا بدی مدل برآورد شده $f(z, \hat{\theta}_n)$ از لحاظ پیش‌بینی برای آینده است که در آن Z به عنوان مشاهده آینده از چگالی h و مستقل از X_i ها است. میزان نزدیکی $f(z, \hat{\theta}_n)$ به توزیع درست مشاهدات از اطلاع KL به صورت

$$
\begin{aligned}
KL\big(h(z), f(z,\hat{\theta}_n)\big) &= E_h\Big[\log\Big(\frac{h(Z)}{f(Z;\hat{\theta}_n)}\Big)\Big] \\
&= E_h\big[\log h(Z)\big] - E_h\big[\log f(Z;\hat{\theta}_n)\big],
\end{aligned}
$$

محاسبه می‌شود در اینجا Z از $\hat{\theta}_n$ مستقل است زیرا تابعی از X_i ها است. با توجه به ویژگی‌های معیار اطلاع KL بزرگ بودن $E_h\big[\log f(Z;\hat{\theta}_n)\big]$ نزدیکی چگالی $f(z;\hat{\theta}_n)$ به چگالی درست h را نشان می‌دهد. چون $E_h\big[\log f(Z;\hat{\theta}_n)\big]$ کمیتی مجهول است باید برآورد شود. یک برآورد از این کمیت را می‌توان با جایگزین کردن تابع توزیع تجربی $\hat{H}$ به جای تابع توزیع درست و مجهول H، به صورت زیر به دست آورد:

[۱۵] Psedo-true Value of a Parameter

$$E_{\hat{H}}\Big[\log f(Z;\hat{\theta}_n)\Big] \quad = \quad \int \log f(Z;\hat{\theta}_n)d\hat{H}(x) = \frac{۱}{n}\sum_{i=۱}^{n}\log f(x_i;\hat{\theta}_n).$$

بنابراین $E_{\hat{H}}\Big[\log f(Z;\hat{\theta}_n)\Big]$ برآوردگری به شکل $\frac{۱}{n}\sum_{i=۱}^{n}\log f(x_i;\hat{\theta}_n)$ دارد.

AIC یک معیار مهم و زیربنایی در انتخاب مدل است. در واقع این معیار از نظر تئوری و عملی مورد استفاده آمارشناسان و تحلیل گران داده‌ها قرار می‌گیرد. AIC یک معیار ارزیابی برای عدم برازش مدل‌هایی است که در آن پارامترها به روش درستنمایی ماکسیمم برآورد شده‌اند و نشان می‌دهد که اریبی لگاریتم درستنمایی متناسب با پارامترهای مدل (p) است. معیار AIC به صورت زیر تعریف می‌شود:

$$AIC = -۲\sum_{i=۱}^{n}\log f(x_i;\hat{\theta}_n) + ۲p.$$

در واقع آکائیک در سال ۱۹۷۴ بیان کرد که اگر چگالی مولد داده‌ها به مدل پارامتری پیشنهاد شده نزدیک باشد میزان اریبی بوسیله تعداد پارامترهای پیشنهاد شده تقریب زده می‌شود. محاسبه معیار AIC به چگالی درست و مجهول h بستگی ندارد.

۱—۷ معیارهای دیگر

در این بخش معیارهای هلینگر و خی دو که در فصل ۲ برای بررسی کران‌های معیار اطلاع کولبک — لیبلر استفاده می‌شود بیان شده است.

تعریف ۱۰.۱ معیار هلینگر[۱۶]

فرض کنید $f(x)$ و $h(x)$ برای هر $x \in \chi$ دو تابع چگالی (احتمال) باشند. آنگاه این معیار به صورت

$$D_H(h,f) = \begin{cases} \frac{۱}{۲}\int_{\chi}\left(\sqrt{h(x)}-\sqrt{f(x)}\right)^{۲}dx, & \text{درحالت پیوسته} \\ \frac{۱}{۲}\sum_{x\in\chi}\left(\sqrt{h(x)}-\sqrt{f(x)}\right)^{۲}. & \text{در حالت گسسته} \end{cases}$$

تعریف می‌شود.

تذکر: معیار هلینگر یک متر نیست اما متقارن است.

۱۶ Hellinger Distance

تعریف ۱۱.۱ معیار خی دو[۱۷]

این معیار که معیاری نامتقارن است به دو صورت تعریف می‌شود.

الف)

$$D_{\chi^{۲}}(h,f)=\begin{cases}\displaystyle\int_{\chi}\frac{h^{۲}(x)-f^{۲}(x)}{f(x)}dx, & \text{درحالت پیوسته}\\ \displaystyle\sum_{x\in\chi}\frac{h^{۲}(x)-f^{۲}(x)}{f(x)}. & \text{در حالت گسسته}\end{cases}$$

ب)

$$D_{\chi^{۲}}(h,f)=\begin{cases}\displaystyle\int_{\chi}\frac{\big(h(x)-f(x)\big)^{۲}}{f(x)}dx, & \text{درحالت پیوسته}\\ \displaystyle\sum_{x\in\chi}\frac{\big(h(x)-f(x)\big)^{۲}}{f(x)}. & \text{در حالت گسسته}\end{cases}$$

اگر تعاریف الف و ب بسط داده شوند رابطه‌های یکسانی بدست می آید.

در این فصل برخی تعاریف مورد نیاز فصل‌های بعد، مدل‌های آماری و معیارهای انتخاب مدل بیان شد. در فصل بعد برخی کران‌های مهم معیار اطلاع کولبک — لیبلر مورد بررسی قرار خواهد گرفت.

---

۱۷ $\chi^{۲}$ *Distance*