آنالیز داده با استفاده از تابع جریمه

درس کیهان‌شناسی

زمستان ۱۴۰۱

تصور کنید که تعدادی نقطه در اختیار دارید و می‌خواهید منحنی‌ای که برازندهٔ آن نقاط است (منحنی برازش) را پیدا کنید. ابتدا باید تکلیفتان مشخص باشد که چه جور منحنی‌ای می‌خواهید فیت کنید به آن نقاط. خط؟ سهمی؟ تانژانت؟ کسینوس؟ لگاریتم؟ تابع بسل؟ ترکیبی از همهٔ اینها؟ هیچ کدام؟

وقتی نوع منحنی را مشخص کردید، باید آزادی‌های منحنی خود را بشناسید. به این معنی که هر منحنی تعدادی پارامتر آزاد دارد. مثلا خط، دو پارامتر آزاد دارد: شیب و عرض از مبدا. ممکن است از جایی مطمئن باشید که خط شما از مبدا عبور می‌کند. پس تنها یک پارامتر آزاد باقی می‌ماند.

بعد از دانستن آزادی‌های منحنی، باید در فضای مقادیر ممکن برای آنها آنقدر بگردید تا بهترین مقدار برای آن پارامترهای آزاد بدست بیاید. بهترین مقدار از نظر چه کسی؟ از نظر داده‌ها!

چگونه خوب بودن یا بد بودن وضعیت یک منحنی از نظر داده‌ها کمّی می‌شود؟ به هزار و یک طریق. یکی از این راه‌ها را اکنون بررسی می‌کنیم.

فرض کنید داده‌های شما به صورت $y = y(x) \pm \delta y$ باشند و شما n داده دارید. پس در واقع nتا سه‌تایی مرتب به صورت $(x_i, y_i, \delta y_i)$ در اختیار شماست. از سوی دیگر فرض کنید منحنی‌ای که تصمیم دارید به این داده‌ها برازش کنید، تابعی با مثلا سه پارامتر آزاد a، b و c باشد. پس اگر مقدار این سه پارامتر مشخص باشد، در هر x تابع شما مقدار مشخصی به خود می‌گیرد: $y = f(x, a, b, c)$.

حالا «تابع جریمه»، که به صورت زیر است را در نظر بگیرید:

$$\chi^2(a,b,c) = \sum_{i=1}^{n} \left(\frac{y_i - f(x_i, a, b, c)}{\delta y_i} \right)^2. \qquad (۱)$$

به خاطر بیاورید که x_i، y_i و δy_i مربوط به داده هستند. تابع χ^2 در واقع معیاری از فاصلهٔ منحنی (یا به عبارتی مدل یا تئوری) تا نقاط داده است. هر چه این تابع کمتر باشد، منحنی به نقاط نزدیک‌تر است (تئوری یا مدل بهتری دارید). جملهٔ iام از این تابع، مشخص می‌کند که منحنی ما چقدر تا دادهٔ iام فاصله(ی عمودی) دارد. هر چه این فاصله بیشتر باشد، مقدار آن جمله هم بالاتر می‌رود و کل χ^2 مقدار بزرگتری به خود می‌گیرد و این یعنی مدل بیشتر جریمه شده است. نقش δy_i در مخرج جملات χ^2 این است که اهمیت یا وزن نقاط داده را مشخص کند: اگر یک داده خطای زیادی داشته باشد، منحنی می‌تواند از آن فاصله بگیرد و در عین حال زیاد جریمه نشود و برعکس، اگر یک داده خطای کمی داشته باشد، فاصله گرفتن منحنی از آن جریمهٔ زیادی در پی خواهد داشت. اگر خطای داده‌ها مشخص نباشد، اهمیت همهٔ داده‌ها را یکسان تلقی کنید و مخرج‌ها را 1 بگذارید.

پس تا اینجا، مدل (نوع منحنی) مشخص است؛ تعدادی داده در اختیار داریم؛ و فقط باید بهترین مقدار پارامترهای آزاد مدلمان (از نظر داده) را پیدا کنیم. چگونه؟ به کمک تابع جریمه: کافیست کمینهٔ این تابع را بیابیم و ببینیم به ازای چه مقادیری از a، b و c تابع جریمه کمینه می‌شود. برای این کار می‌توان به حساب دیفرانسیل

متوسل شد:

$$\frac{\partial \chi^2}{\partial a} = 0, \quad (۲)$$
$$\frac{\partial \chi^2}{\partial b} = 0, \quad (۳)$$
$$\frac{\partial \chi^2}{\partial c} = 0. \quad (۴)$$

سه معادله برای سه مجهول a، b و c و تمام. برای انجام این عملیات (ساختن تابع جریمه با توجه به نقاط داده، مشتق گرفتن از آن و حل معادلات فوق) می‌توانید از نرم‌افزارهایی مثل Mathematica بهره بگیرید.
البته برای پیدا کردن کمینه تابع جریمه و مقادیر پارامترهای آزاد متناظر با آن راه‌های دیگری هم هست: می‌توان کل فضای پارامتری را به صورت گسسته جارو کرد و این‌گونه دید که به ازای کدام مقادیرِ پارامترهای آزاد، تابع جریمه کمینه می‌شود. یا می‌توان به صورت تصادفی (و در عین حال هوشمندانه) در فضای پارامترهای آزاد جست و خیز کرد و کمینهٔ تابع جریمه را پیدا کرد (الگوریتم‌های مونته‌کارلو). اینکه کدام یک از این روش‌ها بهتر است، مورد به مورد متفاوت است.

حالا که بهترین مقدار برای پارامترهای آزاد مدل پیدا شد، سزاوار است آنها را به همراه نایقینی گزارش کنید: $a = a_0{}^{+\delta a_+}_{-\delta a_-}$. یعنی حد بالا و پایین پارامتر آزاد، به شرطی که مدل «زیاد» جریمه نشود، را هم بیان کنید. اینکه «زیاد» یعنی چقدر جریمه، بسته به این است که مدل چند پارامتر آزاد دارد. اگر مدل یک پارامتر آزاد داشته باشد، برای تعیین نایقینی آن پارامتر، $\Delta\chi^2$ را باید برابر با 1 قرار دهید و ببینید پارامتر شما چه مقدار می‌تواند بالا و پایین برود به طوری که تابع جریمه یک واحد زیاد شود. برای دو پارامتر آزاد، $\Delta\chi^2 = 2.3$ و برای مدل‌هایی با سه پارامتر آزاد، $\Delta\chi^2 = 3.53$ و الی آخر. توضیح اینکه این اعداد چطور تعیین شده‌اند در این مقال نمی‌گنجد و می‌توانید برای دریافتن آن به کتاب‌های آمار و مدلسازی مراجعه کنید.