Chapter 4

Multidimensional Scaling

مقیاس بندی چند بعدی

مقدمه

در فصل ۳، به طور گذرا به این نکته اشاره کردیم که یکی از مفیدترین روش های استفاده تجزیه و تحلیل مولفه های اصلی برای به دست آوردن یک \نقشه کم بعدی از آن بود

در این فصل قصد داریم علاوه بر روش تجزیه و تحلیل مؤلفه های اصلی روش صریح تری را معرفی کنیم دسته‌ای از روش‌های دیگر، با برچسب‌گذاری مقیاس‌گذاری چند بعدی، که هدفشان تولید نقشه‌های مشابهی از داده‌ها است، اما مستقیماً روی چند متغیره معمولی کار نمی‌کنند.

مقدمه

در جامعه‌شناسی چندبعدی، یک روش وجود دارد به نام مقیاس‌گذاری چندبعدی که به منظور تولید یک "نقشه" عمل می‌کند، اما به‌طور مستقیم بر روی ماتریس داده‌های چندمتغیره عادی PCA کم‌بعدی از داده‌هاست که مشابه عمل نمی‌کند. به‌جای این کار بر روی ماتریس فواصل اعمال می‌شود که از ماتریس داده‌ها به‌دست می‌آید و همچنین بر روی ماتریس‌های "ماتریس های عدم تشابه" یا "شباهت" که به‌طور مستقیم از روی ارزیابی‌های انجام شده توسط افراد در مورد اینکه چقدر جفت‌های اشیاء، تحریک‌ها و غیره که مورد توجه هستند، شبیه یکدیگر هستند، به‌دست می‌آیند. اصطلاح "مجاورت" اغلب برای شامل شدن هر دو امتیاز عدم تشابه و شباهت استفاده می‌شود. مقیاس‌گذاری چندبعدی در اصل یک تکنیک کاهش داده است زیرا هدف آن پیدا کردن یک مجموعه از نقاط در بُعد کم است که به‌طور تقریبی تنظیمات ممکنه با بُعد بالاتری که توسط ماتریس نزدیکی اصلی نمایش داده می‌شود، را تقریباً تداعی کنند

تجزیه تحلیل مولفه اصلی

روش تجزیه تحلیل مولفه اصلی یا PCA (Principal Component Analysis) یک روش آماری است که برای تحلیل و کاهش ابعاد داده‌های چند متغیره استفاده می‌شود.

که در فصل کامل توضیح داده شده است هدف اصلی PCA این است که با حفظ اطلاعات مهم و حذف اطلاعات تکراری یا غیرضروری، داده‌های پیچیده را به یک فضای کم‌بعدی تبدیل کند. این کار با تبدیل متغیرهای اولیه به یک مجموعه جدید از متغیرهای خطی و مستقل به نام "اجزای اصلی" انجام می‌شود.

مدل‌ها برای داده‌های مجاورت

مدل‌ها به مجاورت ها منطبق می‌شوند تا ساختار یا الگوهای موجود یا محاسبه شده در مجاورت ها که به‌طور آشکار در مجموعه‌ای از ارقام قابل مشاهده نیست، روشن و قابل فهم شود و احتمالاً توضیح داده شود در برخی حوزه‌ها، به‌ویژه روان‌شناسی، هدف نهایی در تحلیل مجموعه‌ای از مجاورت ها مشخص‌تر است، به‌طور خاص توسعه تئوری‌ها برای توضیح دادن داوری‌های شباهت؛ به‌عبارت دیگر، تلاش برای پاسخ به سوال "چه چیزی باعث می‌شود چیزها شبیه یکدیگر به نظر برسند یا متفاوت؟" مدل‌های تحلیل داده‌های مجاورت ها می‌توانند به سه دسته زیر تقسیم شوند: مدل‌های فضایی، مدل‌های درختی و مدل‌های هجی. در این فصل، ما فقط با اولین این سه دسته سر و کار داریم، به مدل‌های فضایی

مدل‌های فضایی برای مجاورت ها:
مقیاس‌گذاری چندبعدی(MDS)

یک نمایش فضایی از یک ماتریس مجاورت شامل مجموعه‌ای از مختصات چند-بعدی است، هر یک از آنها یکی از واحدهای داده را نمایان می‌کند. مختصات مورد نیاز به طور کلی با کمینه کردن برخی اندازه‌گیری از "تطابق" بین فواصلی که توسط مختصات نمایان می‌شوند و مجاورت های مشاهده شده، پیدا می‌شوند. به عبارت ساده، یک مدل هندسی جستجو می‌شود که هرچه فاصله یا عدم تشابه مشاهده شده بین دو واحد بزرگتر باشد (یا شباهت آنها کوچک‌تر)، فاصله بیشتری بین نقاط نمایان دهنده آنها در مدل وجود داشته باشد. به طور کلی،فواصل بین نقاط در مدل فضایی اقلیدسی فرض می‌شود.

یافتن مجموعه بهترین مختصات و مقدار مناسب که برای به‌طور کافی نمایش دادن مجاورت های مشاهده شده لازم است، هدف از روش‌های متعددی از مقیاس‌گذاری چندبعدی است که پیشنهاد شده است. امید این است که تعداد بُعدها، ، کوچک باشد، ایده‌آل دو یا سه، تا تنظیم فضایی مشتق شده به راحتی قابل رسم شود. تنوع روش‌هایی که پیشنهاد شده‌اند اکثراً در این مورد متفاوت است که چگونگی موافقت بین فواصل مناسب شده و مجاورت های مشاهده شده ارزیابی می‌شود. در این فصل، دو روش "مقیاس‌گذاری چندبعدی کلاسیک" و "مقیاس‌گذاری چندبعدی غیرمتریک" را مورد بررسی قرار می‌دهیم

مقیاس بندی چند بعدی کلاسیک

مقیاس‌بندی کلاسیک در واقع سعی دارد یک ماتریس مجاورت را با استفاده از یک مدل یا نقشه هندسی ساده در ابعاد خاص است، به طوری که هر نقطه xn، ...، x2، x1نمایش دهد. این مدل شامل یک مجموعه نقاط MDS ،نمایانگر یکی از واحدهای مورد نظر است و فاصله بین هر جفت از این نقاط را نشان می‌دهد. هدف اصلی است، به گونه‌ای که این xn، ...، x2، x1بعدی nmنمایانده می‌شود و مختصات mتعیین ابعاد مدل که به عنوان مدل تطابق مناسبی با مجاورت‌های مشاهده شده داشته باشد. این تطابق اغلب توسط شاخص‌های عددی ارزیابی می‌شود که در چه اندازه‌ای مجاورت‌ها و فواصل در مدل هندسی هماهنگ هستند. به طور خلاصه، هر چه تفاوت مشاهده شده بین دو واحد بیشتر باشد (یا شباهت آنها کمتر باشد)، نقاط متناظر با آنها در مدل هندسی نهایی باید بیشتر از هم فاصله داشته باشند.

حال سوال اینجاست که چگونه مقدار m و مختصات x_1 ، x_2، ...، x_n از ماتریس مجاورت مشاهده شده برآورد می‌شود. مقیاس‌بندی کلاسیک به این سوال پاسخ می‌دهد. برای شروع، باید توجه داشت که مجموعه‌ای یکتا از مختصات وجود ندارد که باعث ایجاد مجموعه‌ای از فواصل شود، زیرا فواصل تغییری نمی‌کنند با جابجایی کلی پیکربندی نقاط از یک مکان به مکان دیگر یا با چرخش یا بازتاب پیکربندی. به عبارت دیگر، ما نمی‌توانیم به طور یکتا مکان یا جهت نقاط را تعیین کنیم. مشکل مکان معمولاً با قرار دادن میانگین بردار پیکربندی در مبدا حل می‌شود. مسئله جهت یابی به این معناست که هر پیکربندی مشتق شده می‌تواند تحت یک تبدیل متقارن دلخواه قرار گیرد. همانطور که بعداً مشاهده خواهد شد، اغلب از چنین تبدیلاتی برای تسهیل در تفسیر راه‌حل‌ها استفاده می‌شود

مقیاس بندی چند بعدی کلاسیک: جزئیات فنی

فرض کنید که ماتریس مجاورت که با آن سر و کار داریم، یک ماتریس فواصل اقلیدسی با نام Dاست، که از یک ماتریس اولیه n*p به نام X به دست آمده است. در فصل ۱، ما دیدیم چگونه می‌توانیم فواصل اقلیدسی را از ماتریس X محاسبه کنیم؛ اما چندضلعی چند جمله‌ای کلاسیک اساساً با مسئله برعکس سر و کار دارد: با فرض داشتن فواصل، چگونه می‌توانیم ماتریس X را پیدا کنیم؟

ابتدا فرض کنید که X شناخته شده است و ماتریس ضرب داخلی n*n، Bاست.

$$\boldsymbol{B} = \boldsymbol{X}\boldsymbol{X}' \qquad (4.1)$$

المان‌های B به این صورت تعریف می‌شوند:

$$b_{ij} = \sum_{k=1}^{q} x_{ik}x_{jk} \qquad (4.2)$$

به راحتی می توان فهمید که فاصله های اقلیدسی مجذور بین ردیف های X را می توان بر حسب عناصر B به صورت زیر نوشت

$$d_{ij}^2 = b_{ii} + b_{jj} - 2b_{ij} \tag{4.3}$$

اگر بتوان S ها را به عنوان S های موجود در معادله (۴,۳) پیدا کرد، آنگاه مقادیر مختصات مورد نیاز می‌توانند با استفاده از تجزیه به عنوان (۴,۱) بدست آید. راه حلی یکتا وجود ندارد مگر اینکه یک محدودیت مکانی ارائه شود؛ معمولاً مرکز نقاط (X) در مبدا قرار داده می‌شود، به طوری که برای همه k=1,...,m ، این محدودیت‌ها و رابطه‌ای که در (۴,۲) داده شده است، نشان می‌دهد که جمع عبارات در هر ردیف از ماتریس B باید صفر باشد. بنابراین، جمع کردن رابطه‌ای که در (۴,۲) داده شده است بر روی i ، بر روی j ، و در نهایت هم بر روی هر دو i و j منجر به مجموعه‌ای از معادلات می‌شود.

$$\sum_{i=1}^{n} d_{ij}^2 = T + nb_{jj}, \quad \sum_{j=1}^{n} d_{ij}^2 = T + nb_{ii}, \quad \sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2 = 2nT$$

جایگذاری $T = \sum_{i=1}^{n} b_{ii}$ که کلیت ماتریس B است. حالا المان‌های B را می‌توان به صورت فواصل مربعی اقلیدسی یافت.

$$b_{ij} = -\frac{1}{2}\left(d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d_{..}^2\right)$$

$$d_{i.}^2 = \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2, \quad d_{.j}^2 = \frac{1}{n}\sum_{i=1}^{n} d_{ij}^2, \quad d_{..}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2$$

اکنون که عناصر B را بر حسب فواصل اقلیدسی استخراج کرده‌ایم، باید آن را فاکتور کنیم تا مقادیر مختصات را ارائه دهیم. از نظر تجزیه طیفی آن (به فصل ۳ مراجعه کنید)، Bرا می توان به صورت نوشتاری نوشت

$$B = V\Lambda V'$$

هنگامی که در ان $\Lambda = diag(\lambda_1, ..., \lambda n)$ ماتریس قطری مقادیر ویژه B و V ماتریس مربوط به بردارهای ویژه است، به طوری که مجموع مربعات عناصر آنها نرمال شده است. یعنی $v_i v_i' = 1$. فرض می شود که مقادیر ویژه به گونه ای برچسب گذاری می شوند که $\lambda_1 >= \lambda_2 >= ... >= \lambda n$.
وقتی D از یک ماتریس $q * n$ با رتبه کامل ناشی می شود، آنگاه رتبه B برابر q است، به طوری که آخرین n-q از مقادیر ویژه آن صفر خواهد بود. بنابراین B را می توان به صورت نوشت

$$\boldsymbol{B} = \boldsymbol{V}_1 \boldsymbol{\Lambda}_1 \boldsymbol{V}_1'$$

که در آن v_1 شامل اولین بردارهای ویژه q و λ_1 مقادیر ویژه غیر صفر است.
بنابراین مقادیر مختصات به صورت زیر هستند

$$\boldsymbol{X} = \boldsymbol{V}_1 \boldsymbol{\Lambda}_1^{\frac{1}{2}}$$

where $\boldsymbol{\Lambda}^{\frac{1}{2}} = \mathrm{diag}\left(\lambda_1^{\frac{1}{2}}, \dots, \lambda_q^{\frac{1}{2}}\right)$.

استفاده از تمام ابعاد q منجر به بازیابی کامل ماتریس فاصله اقلیدسی اصلی می شود. بهترین برازش نمایش ابعاد m توسط m بردارهای ویژه B مربوط به m بزرگترین مقادیر ویژه داده می شود. کفایت نمایش ابعاد m را می توان با اندازه معیار قضاوت کرد

$$P_m = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{n} \lambda_i}$$

مقادیر Pm از مرتبه ۰,۸ برازش منطقی را نشان می دهد.

در اینجا لازم به ذکر است که در جایی که ماتریس مجاورت حاوی فواصل اقلیدسی محاسبه شده از یک ماتریس داده X n*q است، مقیاس بندی کلاسیک را می توان معادل تجزیه و تحلیل مولفه های اصلی نشان داد، با مقادیر مختصات مورد نیاز مربوط به امتیازات مولفه اصلی استخراج شده است. از ماتریس کوواریانس داده ها. یکی از نتایج این دوگانگی این است که مقیاس بندی چند بعدی کلاسیک به عنوان مختصات اصلی نیز نامیده می شود. و راه حل مولفه های اصلی m بعدی" (m<q) بهترین" است به این معنا که اندازه گیری تناسب را به حداقل می رساند.

$$S = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(d_{ij}^2 - \left(d_{ij}^{(m)} \right)^2 \right)$$

که در آن dij فاصله اقلیدسی بین افراد i و j بر اساس مقادیر متغیر q اصلی آنها است و $d_{ij}^{(m)}$ فاصله متناظری است که از امتیازات مولفه اصلی m محاسبه می شود.
وقتی ماتریس مجاورت مشاهده شده B اقلیدسی نباشد، ماتریس قطعی مثبت نیست. در چنین مواردی، برخی از مقادیر ویژه B منفی خواهند بود. به همین ترتیب، برخی از مقادیر مختصات اعداد مختلط خواهند بود. با این حال، اگر B فقط تعداد کمی از مقادیر ویژه منفی کوچک داشته باشد، نمایش مفید ماتریس مجاورت ممکن است با استفاده از بردارهای ویژه مرتبط با m بزرگترین مقادیر ویژه مثبت.
کفایت راه حل حاصل را می توان با استفاده از یکی از دو معیار زیر ارزیابی کرد

$$P_m^{(1)} = \frac{\sum_{i=1}^{m} |\lambda_i|}{\sum_{i=1}^{n} |\lambda_i|}$$

$$P_m^{(2)} = \frac{\sum_{i=1}^{m} \lambda_i^2}{\sum_{i=1}^{n} \lambda_i^2}$$

دوباره به دنبال مقادیر بالاتر از ۰,۸ برای ادعای تناسب "خوب" هستیم.روش دیگر، دو معیار برای تصمیم گیری در مورد تعداد ابعاد برای مدل فضایی وجود دارد

برای نشان دادن مناسب مجاورت های مشاهده شده:
معیار ردیابی: تعداد مختصات را طوری انتخاب کنید که مجموع مقادیر ویژه مثبت تقریباً برابر با مجموع همه مقادیر ویژه باشد.
معیار بزرگی: فقط آن دسته از مقادیر ویژه را به‌عنوان مثبت واقعی بپذیرید که بزرگی آنها به طور قابل‌توجهی از بزرگترین مقدار ویژه منفی بیشتر باشد.
با این حال، اگر ماتریس B دارای تعداد قابل توجهی مقادیر ویژه منفی بزرگ باشد، مقیاس بندی کلاسیک ماتریس مجاورت ممکن است توصیه نشود و برخی روش های دیگر مقیاس بندی، به عنوان مثال مقیاس بندی غیر متریک (به بخش بعدی مراجعه کنید) بهتر است استفاده شود