پیدا کردن تشکل‌های ژانری در فیلم‌ها با استفاده از پایتون، نُرم‌های گوناگون و ماتریس‌های لاپلاسی

علیرضا سجادی‌پور
۶۱۰۳۹۸۱۳۱

۲۵ آذر ۱۳۹۹

چکیده

وب‌سایت *مووی‌لنز* دیتاستی ۳ـبعدی از امتیازهای ۶۰۰۰ کاربر بر روی ۴۰۰۰ فیلم به طور رایگان در اختیار قرار داده است. هدف از پروژهٔ من، پیدا کردن تشکل‌های (کلاسترها) ژانری در این ۴۰۰۰ فیلم با استفاده از پایتون و روش‌های جبر خطی است. روش اصلی که در ادامه توضیحی مختصر از چگونگی آن داده می‌شود، با استفاده از نرم‌های گوناگون است. اما از کاربرد مقوله‌ای به نام ماتریس‌های لاپلاسی در این زمینه آگاهم و ممکن است از آن‌ها نیز در پروژه‌ام استفاده کنم.

۱ نُرم اقلیدسی، کی به کی شبیه است؟

نرم اقلیدسی که در حالت استاندارد برای فضای $\mathbb{R}^{۲}$ تعریف می‌شود، قابل تعمیم به ابعاد بزرگتر نیز هست. فرض کنید من (علیرضا) و دوستانم فرناز و پارسا بر روی چهار فیلم ***فانوس دریایی*** **(۲۰۱۹)**، ***او*** **(۲۰۱۳)**، ***ممنتو*** **(۲۰۰۰)**، و ***جاذبه*** **(۲۰۱۳)** نظراتی داریم. هر فردی نظرش راجع به فیلم را با یک عدد بین ۵− تا ۵ مشخص می‌کند. نظرات هر فرد را داخل یک بردار ستونی قرار می‌دهیم. مثلاً نظر من $A = \begin{pmatrix} ۵ \\ ۵ \\ ۱ \\ ۰ \end{pmatrix}$ می‌باشد. در جدول زیر می‌توانید نظرات دوستان من بر روی این چهار فیلم را ببینید.

پارسا	**فرناز**	**علیرضا**	شخص/فیلم
۴	۱	۵	***فانوس دریایی*** **(۲۰۱۹)**
۵	۴	۵	***او*** **(۲۰۱۳)**
۳	۳	۱	***ممنتو*** **(۲۰۰۰)**
۱	۵	۰	***جاذبه*** **(۲۰۱۳)**

جدول ۱: جدول علایق من و دوستانم پارسا و فرناز

یک راه محاسبهٔ شباهت بین سلیقهٔ من و سلیقهٔ فرناز، محاسبهٔ نرم اقلیدسی تفاضل دو بردار سلیقهٔ من و سلیقهٔ فرناز است. یعنی

$$\begin{aligned}\|A-F\|_۲ &= \sqrt{(۵-۱)^۲+(۵-۴)^۲+(۱-۳)^۲+(۰-۵)^۲}\\ &= \sqrt{۱۶+۱+۴+۲۵}\\ &= \sqrt{۴۶}\\ &\approx ۶/۷۸.\end{aligned}$$

به همین منوال، می‌توان شباهت سلیقهٔ من و پارسا را محاسبه کرد. برای پارسا داریم

$$\begin{aligned}\|A-P\|_۲ &= \sqrt{(۵-۴)^۲+(۵-۵)^۲+(۱-۳)^۲+(۰-۱)^۲}\\ &= \sqrt{۱+۰+۴+۱}\\ &= \sqrt{۶}\\ &\approx ۲/۴.\end{aligned}$$

می‌توان مشاهده کرد که تحت این نُرم، سلیقهٔ من بیشتر به پارسا شباهت دارد تا به فرناز. اما خوب است آگاه باشیم که ممکن است نرم‌هایی جنبه‌هایی دیگر از این داده‌ها را بها دهند و در نتیجه من را به فرناز شبیه بخوانند. اما این همهٔ ماجرا نیست. در اینجا از شباهت امتیازهای فیلم‌ها استفاده کردیم تا به شباهت اشخاص پی ببریم. در ادامه نشان می‌دهیم که برعکس این امر هم ممکن است.

۲ نرم اقلیدسی، کدام فیلم به کدام فیلم شبیه است؟

بیایید این دفعه از شباهت امتیازدهی افراد در فیلم‌های گوناگون استفاده کنیم تا به شباهت فیلم‌ها پی ببریم. آیا ممکن است همهٔ فیلم‌های شبیه به یک فیلم مشخص را پیدا کنیم؟ بر می‌گردیم به جدول شمارهٔ ۱ و نُرم اقلیدسی تفاضل دو بردار ***فانوس دریایی*** **(۲۰۱۹)** و ***او*** **(۲۰۱۳)** و همچنین نُرم اقلیدسی تفاضل دو بردار ***ممنتو*** **(۲۰۰۰)** و ***جاذبه*** **(۲۰۱۳)** را محاسبه می‌کنیم تا دریابیم کدام یک از این دو جفت فیلم شباهت بیشتری به هم دارند.

بردارهای ردیفی ***فانوس دریایی*** **(۲۰۱۹)** و ***او*** **(۲۰۱۳)** را به ترتیب با L و H مشخص می‌کنیم. داریم

$$\begin{aligned}\|L-H\|_۲ &= \sqrt{(۵-۵)^۲+(۱-۴)^۲+(۴-۵)^۲}\\ &= \sqrt{۰+۹+۱}\\ &= \sqrt{۱۰}\\ &\approx ۳/۱۶.\end{aligned}$$

همچنین، بردارهای ردیفی ***ممنتو*** **(۲۰۰۰)** و ***جاذبه*** **(۲۰۱۳)** را به ترتیب با M و G مشخص می‌کنیم. داریم

$$\begin{aligned}\|M-G\|_۲ &= \sqrt{(۱-۰)^۲+(۳-۵)^۲+(۳-۱)^۲}\\ &= \sqrt{۱+۴+۴}\\ &= \sqrt{۶}\\ &\approx ۲/۴.\end{aligned}$$

این واقعیت که دو نرم از یک مرتبهٔ بزرگی هستند و مقداری نزدیک به هم دارند، اشاره به این حقیقت دارد که این دو جفت فیلم به یک اندازه به هم شبیه هستند. یعنی ***فانوس دریایی*** **(۲۰۱۹)** همانقدر به ***او*** **(۲۰۱۳)** شبیه است که ***ممنتو*** **(۲۰۰۰)** به ***جاذبه*** **(۲۰۱۳)** شبیه است. و حقاً هم که به همین گونه حس می‌شود!

۳ نرم‌های دیگر

همان طور که در طول کاغذ مرتب اشاره شده است، نرم اقلیدسی تنها ابزار ما نیست. ما می‌توانیم به جای مسافت بین بردارها، از زاویهٔ بین آن‌ها استفاده کنیم. به گونه‌ای که منی که از دو فیلم مفروض a و b به شدت تنفر داشته‌ام و به هر دوی آن‌ها ۵–

داده‌ام با کسی که از این دو فیلم به شدت خوشش آمده و به هر دوی آن‌ها ۵ داده است و کسی که در مورد هر دوی آن‌ها ممتنع است و به آن‌ها ۰ داده است، همگی شبیه پنداشته می‌شویم (هر سهٔ ما بر روی یک خط قرار داریم). شاید در نگاه اول خطا به نظر برسد اما با موشکافی در می‌یابیم که در آن حق وجود دارد. احساسات ما سه فرد در گذار از فیلم a به b به یک اندازه تقویت یا تضعیف شده است. این امر نوید وجود خصوصیات نفسانی مشترک را می‌دهد. محاسبات زوایای بین بردارهایی با بعد بیش از ۲ و ۳ طویل است. در این مقاله از آن‌ها پرش کرده و انجام آن را به کامپیوتر می‌سپارم.

۴ ماتریس‌های لاپلاسی

برای نوشتن یک گراف در شمایل یک ماتریس، از مقوله‌ای به نام *ماتریس همسایگی* استفاده می‌شود. *ماتریس همسایگی* یال‌های میان گره‌های گراف را با ۱ و فقدان آن‌ها را با ۰ نمایش می‌دهد. برای مثال ماتریس همسایگی گراف کامل $k_۴$ در زیر نوشته شده است.

$$A = \begin{pmatrix} ۰ & ۱ & ۱ & ۱ \\ ۱ & ۰ & ۱ & ۱ \\ ۱ & ۱ & ۰ & ۱ \\ ۱ & ۱ & ۱ & ۰ \end{pmatrix}$$

برای محاسبهٔ ماتریس لاپلاسی یک گراف، به یک ماتریس دیگر هم نیاز داریم. آن را با D نمایش می‌دهیم. این ماتریس قطری است و درایهٔ d_{ii} آن برابر جمع ردیف iام ماتریس همسایگی (A) است. این ماتریس D را برای $k_۴$ مجاسبه می‌کنیم.

$$D = \begin{pmatrix} ۳ & ۰ & ۰ & ۰ \\ ۰ & ۳ & ۰ & ۰ \\ ۰ & ۰ & ۳ & ۰ \\ ۰ & ۰ & ۰ & ۳ \end{pmatrix}$$

ماتریس لاپلاسی گراف $k_۴$ تفاضل D از A است. آن را با L نشان می‌دهیم و داریم

$$L = A - D = \begin{pmatrix} ۰ & ۱ & ۱ & ۱ \\ ۱ & ۰ & ۱ & ۱ \\ ۱ & ۱ & ۰ & ۱ \\ ۱ & ۱ & ۱ & ۰ \end{pmatrix} - \begin{pmatrix} ۳ & ۰ & ۰ & ۰ \\ ۰ & ۳ & ۰ & ۰ \\ ۰ & ۰ & ۳ & ۰ \\ ۰ & ۰ & ۰ & ۳ \end{pmatrix} = \begin{pmatrix} -۳ & ۱ & ۱ & ۱ \\ ۱ & -۳ & ۱ & ۱ \\ ۱ & ۱ & -۳ & ۱ \\ ۱ & ۱ & ۱ & -۳ \end{pmatrix}.$$

پیدا کردن نقش این ماتریس در مسائل تشکل‌یابی از کارهای *میُرسلاو فیدلر* است. او کشف کرد که بردار ویژهٔ متناظر با دومین مقدار ویژهٔ کوچک ماتریس لاپلاسی از اهمیت بسیاری برخوردار است. از طریق این بردار می‌توان یک گراف را به زیرـ گراف‌های متصل ماکسیمال و متصل مینیمال تجزیه کرد. نسخه‌های پیچیدهٔ همین الگوریتم امروزه در فیسبوک و اینستاگرم استفاده می‌شوند تا از درون یک جامعه، تشکل‌ها و گروهک‌هایی از دوستان و خانواده‌ها را پیدا کنند. تبلور جبر خطی همین است. فرمولی بسته برای پیدا کردن زیرـ گراف‌های متصل ماکسیمال واقعاً دستاورد بزرگی است. اگر روش‌های جبر خطی نبود، برای پیدایش این تشکل‌ها نیاز به نوشتن الگوریتم‌های باز و بسیار پیچیده در هزاران خط و با گزاره‌های شرطی بسیار بود. اما جبر خطی ناجی پردازنده‌های ما است.

مراجع

[۱] When Life Is Linear: From Computer Graphics to Bracketology - Timothy P. Chartier - 2015
[۲] English Wikipedia