آنالیز مولفه اصلی - PCA

تمرین درس داده کاوی بخش پنجم – استاد مهردوست

کیان رضایی

مقدمه

- آنالیز مولفه اصلی (principal component analysis) تکنیکی چند متغیره میباشد که هدف اصلی آن کاهش بعد (کاهش تعداد متغیرها) یک مجموعه داده ی چند متغیره است تا آن جایی که تا حد ممکن تغییرات متغیرهای اولیه در مجموعه داده را توضیح دهد. این هدف به وسیله تبدیل متغیرهای اولیه به یک مجموعه ی جدید از متغیرهای ناهمبسته با نام مولفه های اصلی به دست می آید که ترکیبات خطی از متغیرهای اصلی (اولیه) هستند و طوری مرتب شده اند که چند مولفه اول، بیشترین تغییر پذیری در متغیرهای اصلی را محاسبه میکنند.

PCA چیست؟

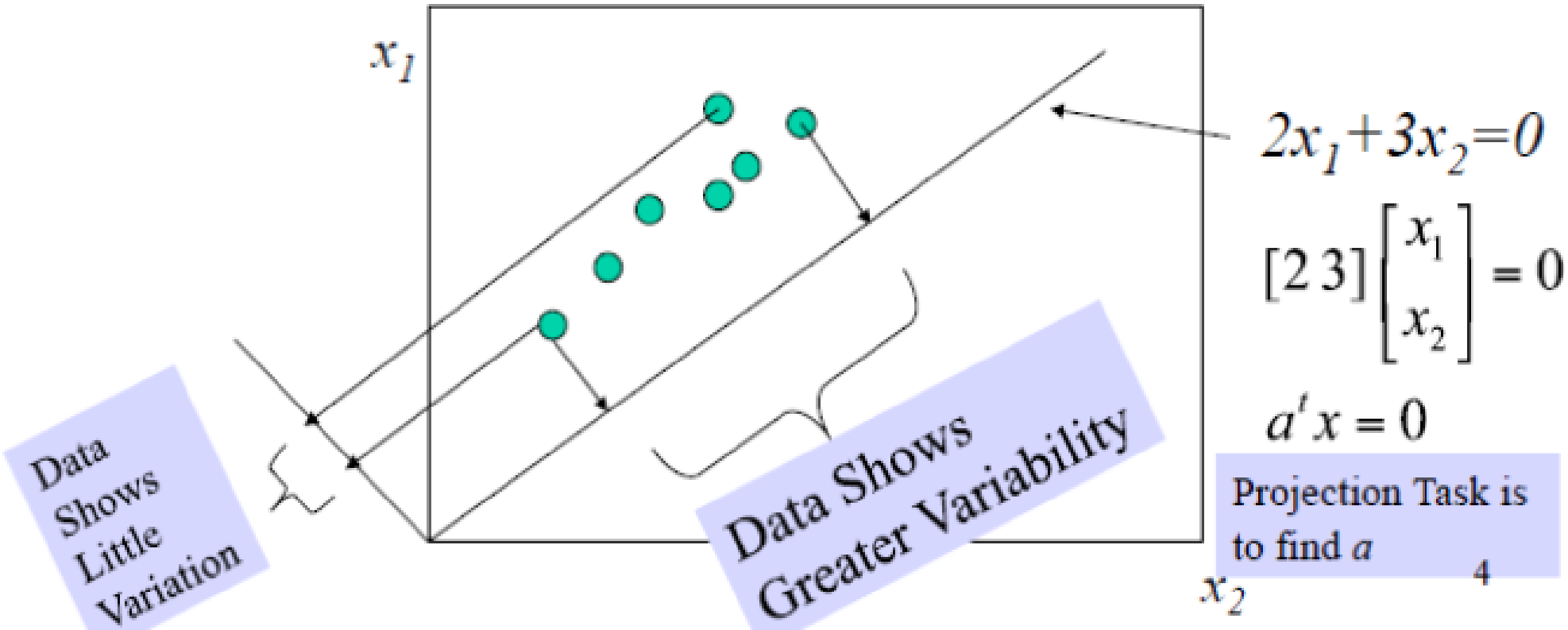
- به اختصار تکنیکی است برای تبدیل یک مجموعه از متغیرهای همبسته مشاهده شده، به منظور توصیف تغییرات ، به یک مجموعه ی جدید از متغییرهایی که با یکدیگر ناهمبسته هستند.
- مزیت و کاربرد اصلی آن کاهش بعد است.

PCA چیست؟

- **انگیزه استفاده از PCA :**

نمودار پراکنش (scatter plot) برای تحلیل روابط بین دو متغیر مناسب میباشند اما ضعف آن ها در روابط پیچیده تر ناشی از تعداد متغیرهای زیاد است. اما PCA با تصویر متغیرها در جهات دیگر سعی میکند روابط بین آن ها را کشف کند. به طور کلی روش های جستجوی تصویر این امکان را به وجود می آورد که جهات مناسب را پیدا کرده و ابعاد را کاهش داد. در شکل زیر دو بعد به تک بعد تبدیل شده...

PCA چیست؟


x_1
x_2
$2x_1+3x_2=0$
$[2\ 3]\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$
$a^t x = 0$
Data Shows Little Variation
Data Shows Greater Variability
Projection Task is to find a

کاربرد های PCA

- استفاده از مولفه های اصلی در حالات زیر ممکن است مفید باشد:

۱. تعداد متغیرهای توضیحی در مقایسه با تعداد مشاهدات بسیار زیاد است.
۲. متغیر های توضیحی به شدت همبسته هستند.

کاربرد های PCA

- به عنوان کاربردی از آنالیز مولفه های اصلی، در حوزه اقتصاد میباشد که داده های پیچیده اغلب به وسیله ی چند نوع عدد شاخص به طور مثال شاخص قیمت، میزان دستمزد، هزینه زندگی و .... خلاصه شده اند. هر گاه مالیات قیمت ها در طول زمان تغیر میکند، اقتصاددان مایل است این واقعیت که قیمت های برخی کالاها تغییر پذیرتر از بقیه بوده یا انیکه قیمت های برخی کالاها مهمتر از بقیه در نظر گرفته شده اند، را بررسی کند. در هر مورد لازم است که شاخص آن موزون شده باشد. در چنین مثال هایی، اغلب م.لفه ی اصلی اول میتواند نیازهای محقق را برطرف کند.

کاربرد های PCA - ادامه

- اما همیشه اولین مولفه ی اصلی برای یک محقق بیشترین نفع را ندارد. به طـور مثـال یـک زیست شناس برای بررسی میزان تغییر در اندازه های ساختار شناسی روی حیوانات چون تمام همبستگی های دوتایی به احتمال زیاد مثبت میباشند. اغلب با مولفه های دوم و بعدی کار میکنند. چون این مولفه ها ممکن است توصیفی مناسب از جنبه هـای مختلـف شـکل حیوان ارائه دهند. دومین مولفه ی اصلی اغلب از جهات اندازه ی حیوان، مـورد علاقـه ی محقق می باشد. که در این جا به دلیل همبستگی های مثبت، روی اولین مولفه های اصـلی تاثیر خواهد گذاشت.

کاربرد های PCA - ادامه

- به عنوان یک مثال دیگر :

اولین مولفه ی اصلی در مقادیر روانشناسی بالینی در بیماران ممکن است تنها شاخص قطع علایم بیماری را فراهم کند و بقیه مولفه ها هستند که اطلاعات مهم روانشناس در مورد الگوی علایم بیماری را خواهند داد.

آنالیز مولفه های اصلی

- هدف اصلی آنالیز مولفه های اصلی، توصیف تغییرات در یک مجموعه از متغیرهای همبسته $x' = (x_1, \dots, x_q)$ برحسب مجموعه ای جدید از متغیرهای ناهمبسته $y' = (y_1, \dots, y_q)$

که هر کدام ترکیبی خطی از متغیرهای x هستند، میباشند. متغیرهای جدید به ترتیب کاهش مرتبه ی اهمیت در نظر گرفته میشوند. بدین ترتیب که y_1 تا حد ممکن بیشترین میزان تغییر در داده های اولیه را در میان تمام ترکیبات خطی x محاسبه میکند. سپس y_2 برای محاسبه ی تا حد ممکن میزان تغییر باقی مانده انتخاب شده به طوری که با y_1 ناهمبسته باشد. متغیر های جدید تعریف شده با این فرایند یعنی

$$y_1, \dots, y_q$$

مولفه های اصلی هستند.

آنالیز مولفه های اصلی

- به طور کلی از تحلیل مولفه های اصلی انتظار می رود که تعداد کمی از اولین مولفه ها، نسبت بزرگی از میزان تغییر در متغیرهای اولیه $x_1, \ldots, x_q$ را محاسبه کنند (توضیح دهند) در نتیجه برای تهیه خلاصه ای مناسب با بعد کمتر از این متغیرها به دلایل گوناگون استفاده میشود.

توضیحات گام به گام PCA

- قدم اول : استاندارد سازی

هدف اصلی این قدم استاندارد سازی دامنه متغیر های اولیه پیوسته است که هر کدام به میزان مساوی در آنالیز سهم داشته باشند.

استاندارد سازی : در آمار استاندارد سازی، به فرایندی گفته میشود که در آن متغیر های مختلف در یک scale قرار داده میشوند.

توضیحات گام به گام PCA – استاندارد سازی

- به طور دقیق تر، دلیل اصلی آنکه استاندارد سازی ضروری است، در نظر گرفتن میزان حساسیت متغیرهای بعدی نسبت به واریانس متغیرهای آغازین است.
- به همین دلیل است، اگر تفاوت زیادی بین دامنه متغیر های آغازین وجود داشته باشد، آن متغیر هایی که دامنه بزرگتری دارند بر آنهایی که دامنه کوچکتری دارند ارجحیت پیدا میکنند. به عنوان مثال یک متغیر که دامنه آن بین ۰ تا ۱۰۰ است بر یک متغیر یک دامنه آن ۰ تا ۱۰ است ارجحیت (dominate) پیدا میکند.) که ما را به سمت نتایج جانبدارانه (biased results) می برد.

توضیحات گام به گام PCA – استاندارد سازی

- بنابراین تبدیل داده به مقیاس های قابل مقایسه میتواند از این مشکل جلوگیری کند.
- از نظر ریاضیاتی، این کار با تفریق کردن از میانگین و تقسیم آن بر انحراف معیار برای هر مقدار از هر متغیر به دست می آید.

$$z = \frac{value - mean}{standard\ deviation}$$

توضیحات گام به گام PCA – محاسبه ماتریس کوواریانس

- دومین قدم : محاسبه ماتریس کوواریانس
- هدف از این قدم فهمیدن اینکه چقدر متغیرهای مجموعه داده ورودی از میانگین دور (متفاوت) هستند. به عبارت دیگر بفهمیم آیا رابطه ای میان آنها برقرار است یا خیر؟
- به این دلیل که گاهی، متغیرها به شدت همبسته هستند و مقادیر افزونه داده زیادی در خود دارند.
- پس به همین دلیل برای شناسایی این همبستگی از ماتریس کوواریانس استفاده و آن را محاسبه میکنیم.

توضیحات گام به گام PCA – محاسبه ماتریس کوواریانس

- برای مثال برای یک دیتاست سه بعدی، با سه متغیر x, y, z ماتریس کوواریانس به فرم زیر است :

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

Covariance Matrix for 3-Dimensional Data

توضیحات گام به گام PCA – محاسبه ماتریس کوواریانس

- از آنجایی که کوواریانس یک متغیر با خودش برابر مقدار واریانس آن متغیر است.

$$cov(a, a) = var(a)$$

پس در واقع ما در قطر اصلی مقدار واریانس متغیر ها را خواهیم داشت.

- از طرفی چون کوورایانس دارای خاصیت جابجایی است یعنی :
- $cov(a, b) = cov(b, a)$

پس ماتریس کوواریانس، ماتریس بالا مثلثی و ماتریس پایین مثلثی یکسانی دارند.

توضیحات گام به گام PCA – محاسبه ماتریس کوواریانس

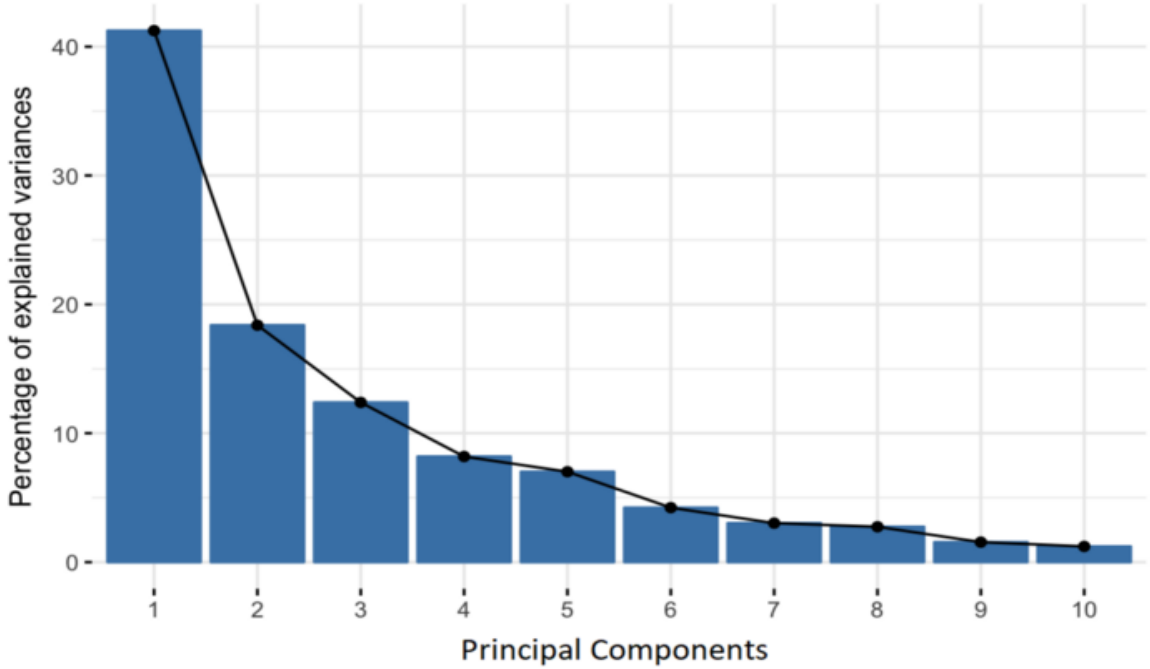
- اما آیا ماتریس کوواریانس همبستگی بین متغیرها را به ما میدهد؟
- در واقع علامت کوواریانس است که بسیار اهمیت دارد زیرا :

۱. **اگر مثبت بود آنگاه : دو متغیر با هم کم یا زیاد میشوند (همبسته هستند)**
۲. **اگر منفی بود آنگاه : یک متغیر زیاد و دیگری کم یا برعکس (غیر همبسته هستند)**

توضیحات گام به گام PCA – بردارهای ویژه و مقدارهای ویژه

- قدم سوم : محاسبه بردار ویژه و مقدار ویژه از ماتریس کوواریانس برای شناسایی مولفه های اصلی است.

مولفه های اصلی : متغیرهای جدید هستند که از ترکیب خطی از متغیرهای اولیه ساخته میشوند.



توضیحات گام به گام PCA – بردار های ویژه و مقدار های ویژه

- طبقه بندی اطلاعات در مولفه های اصلی به این شکل، این امکان را به ما میدهد بدون از دست دادن اطلاعات زیاد بعد را کاهش دهیم.
- نکته مهم در اینجا این است که ، مولفه های اصلی تفسیر خاصی و معنی خاصی ندارند و فقط از ترکیبات خطی متغیر های آغازین به دست می آیند.
- از نظر هندسی، مولفه های اصلی مسیر داده ها که ماکسیمم مقدار واریانس را توضیح میدهد، نشان میدهد. و در واقع خط در شکل بیشترین حجم داده را در خود تسخیر میکند.
- برای سادگی ، به مولفه های اصلی به عنوان یک محور جدید که بهترین زاویه برای دیدن و ارزیابی داده است نگاه کنید.

توضیحات گام به گام PCA – بردارهای ویژه و مقدارهای ویژه

- مثال : فرض کنید که دیتاست ما دو بعدی و با دو متغیر x, y است. که بردارهای ویژه و مقدار های ویژه به شرح زیر باشد :

$$v1 = \begin{bmatrix} 0.6778736 \\ 0.7351785 \end{bmatrix} \qquad \lambda_1 = 1.284028$$

$$v2 = \begin{bmatrix} -0.7351785 \\ 0.6778736 \end{bmatrix} \qquad \lambda_2 = 0.04908323$$

توضیحات گام به گام PCA – بردار های ویژه و مقدار های ویژه - ادامه مثال

- اگر مقادیر ویژه را به صورت نزولی فرض کنیم، داریم : $\lambda_1 > \lambda_2$
- که بدین معنی است اولین مولفه اصلی (PCA۱) همان v_1 است و دومین مولفه اصلی (PCA۲) همان v_2 است.
- بعد از اینکه مولفه های اصلی خود را شناختیم، برای محاسبه درصد واریانس به وسیله هر مولفه، ما مقدار ویژه هر مولفه را تقسیم بر مجموع مقدار ویژه ها میکنیم. اگر این اعمال را برای مثال صفحه قبل انجام دهیم به دست می آوریم که به ترتیب PCA۱ و PCA۲ شامل ۹۶٪ و ۴٪ واریانس داده ها میباشند.

توضیحات گام به گام PCA – بردار ویژگی

- قدم چهارم : بردار ویژگی
- همانطور که در گام قبل دیدیم ، محاسبه بردار ویژه و مرتب کردن آنها بر اساس مقادیر ویژه که نزولی بودند، این امکان را برای ما فراهم کرد که مولفه های اصلی را بسازیم.
- در این گام، کاری که ما انجام میدهیم این است که آیا همه مولفه های اصلی را نگه داریم یا بی ارزش ترین مولفه اصلی را کنار بگذاریم و ماتریس خود را با استفاده از بردار های باقی مانده بسازیم که به آن بردار ویژگی میگویند.
- پس بردار ویژگی یک ماتریس است که ستون های آن شامل بردار های ویژه ای از مولفه ها است که ما تصمیم میگیریم آنها را نگه داریم. این اولین گام در کاهش بعد داده است، زیرا ما تصمیم میگیریم که از n بردار ویژه مولفه ها p تا را انتخاب کنیم. دیتاست نهایی ما در آخر تنها p بعد خواهد داشت!!!

توضیحات گام به گام PCA – بازسازی مجدد داده

- گام آخر : بازسازی مجدد داده بر روی محور مولفه های اصلی است.
- در این گام، که آخرین مرحله است، هدف آن است که بردار ویژگی ها استفاده کنیم تا داده ها از محور اصلی خود به آن محوری که مولفه های اصلی نمایش میدهند جهت گیری مجدد(reorient) نماید.
- فرمول مجموعه داده نهایی به شرح زیر است :

$$Final\ Data\ set = FeatureVector^T * StandardizedOriginalDataset^T$$

منابع

- principal_components.pdf (otago.ac.nz)
- pca.dvi (cmu.edu)
- Python PCA (Principal Component Analysis) with Sklearn – DataCamp
- A Step-by-Step Explanation of Principal Component Analysis (PCA) | Built In