"بسمه تعالی"

گزارش پروژه تحلیل مولفه های اصلی

+

تحلیل نتایج

مفهوم PCA :

تحلیل مولفه اساسی به بیان ساده، روشی برای استخراج متغیرهای مهم (به شکل مولفه) از مجموعه بزرگی متغیرهای موجود در یک مجموعه داده است. تحلیل مولفه اساسی در واقع یک مجموعه با بُعد پایین از ویژگی‌ها را از یک مجموعه دارای بُعد بالا استخراج می‌کند تا به ثبت اطلاعات بیشتر با تعداد کمتری از متغیرها کمک کند. بدین شکل، بصری‌سازی داده‌ها نیز معنادارتر می‌شود. تحلیل مولفه اساسی هنگامی که با داده‌های دارای سه یا تعداد بیشتری بُعد سروکار داشته باشید، کاربردپذیرتر است. تحلیل مولفه اساسی همیشه روی ماتریس کوواریانس یا همبستگی اعمال می‌شود. این یعنی داده‌ها باید عددی و استاندارد شده باشند. برای درک بهتر این روش، در ادامه یک مثال بیان شده است:

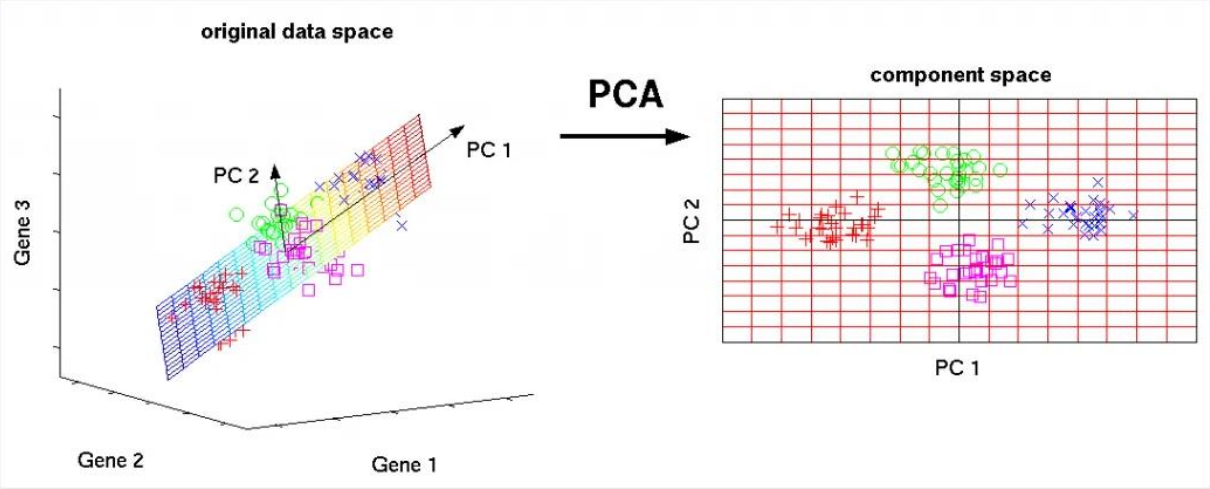
فرض بر آن است که یک مجموعه داده با ابعادی که در زیر آمده است وجود دارد.

50 (p) ×300 (n)

در این مجموعه داده n تعداد کل نمونه‌ها و p تعداد پیش‌بین‌ها (متغیرهای پیش‌بینی) است. به دلیل آنکه تعداد ابعاد زیاد و برابر ۵۰ است می‌توان p(p-1)/2 نمودار پراکندگی برای آن رسم کرد، این یعنی بیش از ۱۰۰۰ نمودار برای انجام تحلیل روی روابط بین متغیرها وجود دارد و در نتیجه تحلیل آن‌ها کاری بسیار خسته‌کننده، دشوار و پیچیده خواهد بود.

در این شرایط، یک رویکرد صحیح می‌تواند آن باشد که یک زیر مجموعه از پیش‌بین‌ها که حاوی بیشترین اطلاعات درباره داده‌ها هستند، انتخاب شود. این امر موجب می‌شود نمودار پراکندگی داده‌ها در ابعاد پایین‌تری قابل ترسیم باشد. تصویر زیر نگاشت داده‌های دارای ابعاد بالا (۳ بُعد) را به داده‌های

با ابعاد پایین‌تر (۲ بُعد) با استفاده از روش تحلیل مولفه اساسی نشان می‌دهد. لازم به ذکر است هر بُعد حاصل شده در فضای جدید، یک ترکیب خطی از p ویژگی اصلی است.



شکل ۱: کاهش ابعاد داده‌ها با استفاده از روش تحلیل مولفه اساسی

یک مولفه اساسی یک ترکیب خطی نرمال شده از پیش‌بین‌های اصلی موجود در مجموعه داده است. در شکل ۱، PC1و PC2 مولفه‌های اساسی هستند. فرض می‌شود یک مجموعه از پیش‌بین‌ها به صورت X^1, X^2...,X^pوجود دارد. مولفه‌های اساسی این مجموعه از پیش‌بین‌ها را می‌توان بدین شکل نوشت:

$$Z^1 = \Phi^{11}X^1 + \Phi^{21}X^2 + \Phi^{31}X^3 + ... + \Phi^{p1}X^p$$

که در آن:

- Z^1اولین مولفه اساسی است.
- Φp^1بردار بار شامل بردارهای بار (Φ^1, Φ^2...) اولین مولفه اساسی است. بردارهای بار به مجموع مربعات مساوی با یک محدود شده‌اند. دلیل این امر آن است که داشتن مقادیر بار بزرگ ممکن است منجر به ایجاد واریانس بسیار بزرگ شود. این مقدار همچنین جهت مولفه اساسی (Z^1) را در جهتی که داده‌ها بیشترین تنوع را دارند، تعریف می‌کند. نتیجه این امر یک خط در فضای

، pبعدی است که نزدیک‌ترین مقدار به n نمونه را دارد. میزان نزدیکی به وسیله محاسبه میانگین مربعات فاصله‌های اقلیدسی اندازه‌گیری می‌شود.

- **X^1..Xpپیش‌بین‌های نرمال شده هستند. میانگین پیش‌بین‌های نرمال شده برابر با صفر و انحراف معیار آن‌ها برابر با یک است.**

بنابراین:

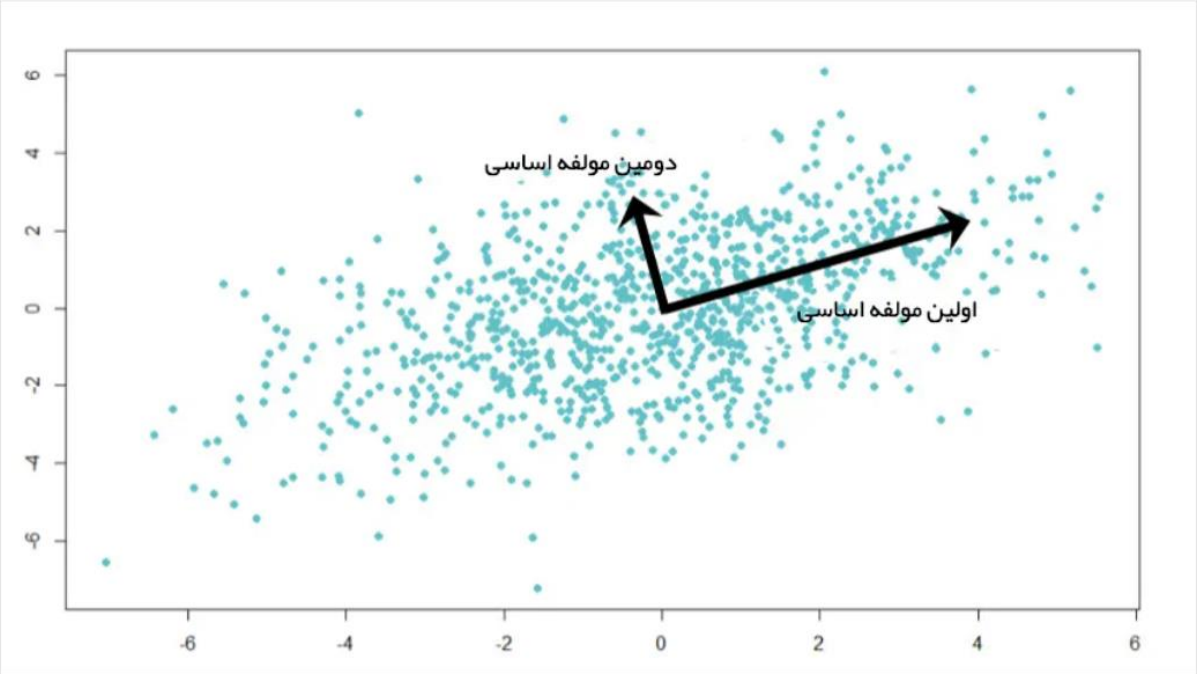
اولین مولفه اساسی، یک ترکیب خطی از پیش‌بین‌های اصلی است که بیشترین واریانس موجود در مجموعه داده‌ها را در بر می‌گیرد. این مولفه، جهت بیشترین تغییرات در داده‌ها را تعیین می‌کند. هرچه دامنه تغییرات موجود در اولین مولفه بالاتر باشد، اطلاعات موجود در این مولفه بیشتر است. هیچ مولفه دیگری نمی‌تواند بیش از مولفه اساسی اول دامنه تغییرات داشته باشد. نتیجه محاسبه اولین مولفه اساسی، خطی است که نزدیک‌ترین خط به داده‌ها محسوب می‌شود. در واقع این خط مجموع مربع فواصل را بین یک نقطه داده و خط، به کمینه مقدار می‌رساند.

مولفه اساسی دوم را نیز به روش مشابهی می‌توان به دست آورد:

دومین مولفه اساسی (Z^2)نیز یک ترکیب خطی از پیش‌بین‌های اصلی است که واریانس باقی‌مانده در مجموعه داده را در خود حفظ می‌کند و با مقدار Z^1 ناهمبسته است. به عبارت دیگر، همبستگی بین مولفه اساسی اول و دوم صفر است. مولفه اساسی دوم را می‌توان به شکل زیر نمایش داد:

$$Z^2 = \Phi^{12}X^1 + \Phi^{22}X^2 + \Phi^{32}X^3 + ... + \Phi p2Xp$$

اگر دو مولفه ناهمبسته باشند، جهت‌های آن‌ها باید متعامد (مانند شکل ۲) باشد. شکل ۲ براساس داده‌های شبیه‌سازی شده با دو ویژگی ترسیم شده است. جهت مولفه‌ها، چنان‌که انتظار می‌رود به صورت متعامد است و این یعنی مقدار همبستگی آن‌ها برابر با صفر است.



شکل ۲: همبستگی مولفه اساسی اول و دوم برابر با صفر و بنابراین بردارهای آن‌ها متعامد است.

کلیه مولفه‌های اساسی بعدی نیز از مفهومی مشابه آنچه بیان شد، پیروی می‌کنند. به عبارت دیگر، آن‌ها مقدار واریانس باقیمانده را بدون آنکه با مولفه‌های پیشین دارای همبستگی شوند، در خود حفظ می‌کنند. به‌طور کلی، در داده‌های دارای $n \times p$ بُعد، به میزان $\min(n-1, p)$ مولفه اساسی قابل ایجاد است.

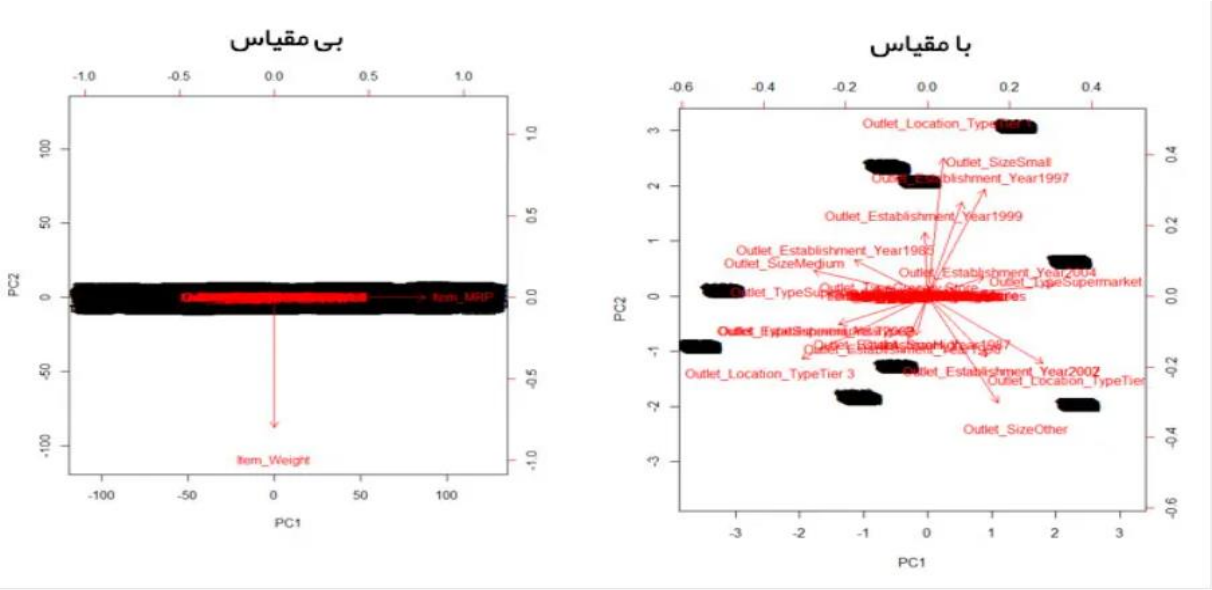
جهت این مولفه‌ها به صورت نظارت نشده تعیین می‌شوند. یعنی، متغیر پاسخ (Y) برای تعیین جهت مولفه استفاده نمی‌شود. بنابراین، این رویکرد نظارت نشده است.

نکته: حداقل مربعات جزئی (PLS) یک جایگزین نظارت شده برای تحلیل مولفه اساسی (PCA) است . PLSبرای تعیین مولفه اساسی، وزن بیشتری را به متغیرهایی که به شدت به متغیر پاسخ مرتبط هستند اختصاص می‌دهد.

علت الزام برای نرمالسازی داده ها :

تحلیل مولفه اساسی روی نسخه نرمال شده پیش‌بین‌های اصلی قابل انجام است. این امر به آن دلیل است که پیش‌بین‌های اصلی ممکن است مقیاس‌های گوناگونی داشته باشند. به عنوان مثال می‌توان به یک مجموعه داده که شامل متغیرهایی با یکاهای گالون، کیلومتر، سال نوری و دیگر انواع واحدها است، اشاره کرد. واضح است که مقدار واریانس این متغیرها اعداد بزرگی خواهد بود. انجام PCA روی متغیرهای نرمال نشده منجر به بارهای فوق‌العاده بزرگی برای متغیرهای دارای واریانس بالا می‌شود و این امر به نوبه خود می‌تواند منجر به وابستگی مولفه اساسی به متغیرهای دارای واریانس بالا شود که بسیار نامطلوب است.

چنانکه در شکل ۳ می‌توان دید، PCAدو بار روی مجموعه داده اجرا گشته (با متغیرهای نرمال شده و نرمال نشده). مجموعه داده به‌کار برده شده در این مثال دارای ۴۰ ویژگی است. چنانکه مشهود است، اولین مولفه اساسی تحت سیطره متغیر MRP قرار گرفته است. دومین مولفه اساسی نیز تحت تسلط متغیر Item_Weight قرار گرفته است. این اتفاقات به دلیل بالا بودن واریانس متغیر است. هنگامی‌که متغیرها نرمال می‌شوند، بصری‌سازی آن‌ها در فضای دو بُعدی به شکل بهتری انجام‌پذیر است.



شکل ۳: تحلیل مولفه اساسی با نرمال‌سازی متغیرها و بدون نرمال‌سازی آن‌ها

تحلیل نتایج :

با در نظر گرفتن دو روش سنتی (مقاله) و بهینه سازی شده (کتابخانه Sk-Learn) مشاهده میشود که علاوه بر تاثیرات خود الگوریتم بر یادگیری مدل ، هایپر پارامتر هایی وجود دارند که دستکاری آنان باعث تغییر دقت مدل میشود. با استفاده از کتابخانه مذکور (که عدد گذاری روی هایپر پارامتر ها به روش کاملا بهینه انجام میشود.) مشاهده شد که دقت مدل افزایش چشم گیری داشته است. موارد دیگری نیز از قبیل نرمالسازی داده روی آموزش و تست داده ها ، تاثیرات مثبتی داشته اند.