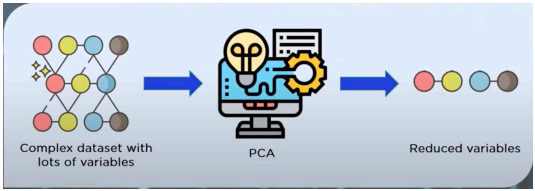
یادگیری بدون نظارت
کاهش ابعاد

المپیاد هوش مصنوعی - یادگیری ماشین

معرفی PCA

PCA که مخفف Principal Component Analysis (تجزیه مولفه‌های اصلی) است، یک روش آماری است که به ما کمک می‌کند اطلاعات زیاد و پیچیده را ساده‌تر کنیم، بدون اینکه بخش مهمی از اطلاعات رو از دست بدیم. انگار که داریم از یک نمای شلوغ، یک خلاصه مفید و مرتب درست می‌کنیم.



معرفی PCA - ادامه

فرض کنید شما مسئول شده‌اید از کلاس درس یک عکس دسته‌جمعی بگیرید. کلاس پر از دانش‌آموز است و همه در جای خود نشسته‌اند. اگر همین‌طوری الکی یک گوشه‌ای بایستید و عکس بگیرید، احتمالاً تعدادی از دانش‌آموزان پشت سر بقیه قرار می‌گیرند و چهره‌شان معلوم نمی‌شود. یا شاید از یک زاویه‌ای عکس بگیرید که همه در یک ردیف قرار گرفته‌اند و عکس حالت تخت و بی‌روحی پیدا می‌کند و اطلاعات کمی از تنوع افراد در اختیارتان می‌گذارد.

معرفی PCA - ادامه

هدف شما چیست؟ هدف شما این است که بهترین زاویه را برای عکس گرفتن پیدا کنید؛ زاویه‌ای که باعث شود بیشترین تعداد چهره‌ها در عکس به وضوح دیده شوند و اطلاعات مهم مربوط به دانش‌آموزان (یعنی چهره‌هایشان که آن‌ها را از هم متمایز می‌کند) تا جای ممکن در عکس ثبت شود. در واقع، می‌خواهید از آن سمتی عکس بگیرید که دانش‌آموزان نسبت به زاویه دید شما بیشترین پخش شدگی را داشته باشند و روی هم نیفتند.

معرفی PCA - ادامه

حالا PCA این کار را در دنیای داده‌ها چگونه انجام می‌دهد؟

- دانش‌آموزان = داده‌های شما: هر دانش‌آموز در کلاس مثل یک "نقطه" یا "نمونه" در مجموعه داده‌های شماست.
- ویژگی‌های دانش‌آموزان (قد، وزن، نمره‌ها و...) یا موقعیتشان در کلاس = متغیرهای شما: اینها همان اطلاعات یا "ابعاد"ی هستند که در مورد هر دانش‌آموز دارید. اگر اطلاعات خیلی زیادی از هر نفر داشته باشید (مثل اینکه علاوه بر قد و وزن، رنگ مو، رنگ چشم، سایز کفش، تعداد خواهر برادر و... را هم بدانید)، انگار هر دانش‌آموز در یک "فضای چند بُعدی" قرار گرفته.

معرفی PCA - ادامه

- زاویه عکس گرفتن = جهت نگاه کردن به داده‌ها: شما می‌توانید از هر جهتی به داده‌هایتان نگاه کنید. اما همه جهت‌ها مفید نیستند.
- پیدا کردن بهترین زاویه عکس گرفتن = پیدا کردن "مولفه‌های اصلی" : PCA می‌آید و دقیقاً همان کاری را می‌کند که شما موقع عکس گرفتن دنبالش هستید: بهترین "زاویه دید" یا "جهت" را در فضای پیچیده داده‌ها پیدا می‌کند؛ جهتی که وقتی داده‌ها را از آن زاویه "نگاه" می‌کنید (یا به زبان ریاضی، روی آن جهت "تصویر" می‌کنید)، داده‌ها بیشترین پخش‌شدگی و تنوع را دارند. یعنی اطلاعات مهمشان (مثل چهره‌های واضح دانش‌آموزان در عکس) روی هم نیفتاده و مشخص است. به این جهت‌ها "مولفه‌های اصلی" می‌گویند.

مثالی از PCA

حالا که ایده کلی PCA را با مثال عکس گرفتن از کلاس فهمیدیم، بیایید با یک مثال ساده‌تر، ببینیم این "پیدا کردن بهترین زاویه" چطور مرحله به مرحله انجام می‌شود.

- این جدول، نقطه شروع ما برای یادگیری گام به گام PCA است.

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

مثالی از PCA - ادامه

در این جدول، ما یک مثال خیلی ساده از داده‌ها را می‌بینیم:

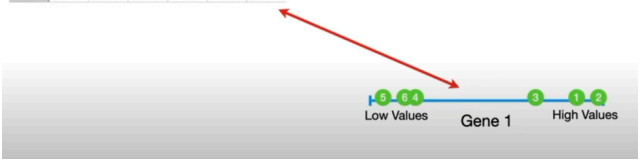
- در سطر اول، اندازه‌گیری‌های مربوط به "ژن ۱" برای ۶ موش مختلف (از موش ۱ تا موش ۶) ثبت شده است.
- در سطر دوم، اندازه‌گیری‌های مربوط به "ژن ۲" برای همان ۶ موش ثبت شده است.

پس، ما اینجا یک مجموعه داده کوچک داریم که شامل اطلاعات دو ویژگی (ژن ۱ و ژن ۲) برای شش نمونه (موش‌ها) است.
این جدول ساده، همان ماده خام اولیه ماست که در اسلایدهای بعدی، از آن استفاده می‌کنیم تا مرحله به مرحله ببینیم PCA چطور این داده‌ها را تحلیل می‌کند، "مهم‌ترین جهت‌ها" یا همان مولفه‌های اصلی را پیدا می‌کند و در نهایت اطلاعات را برایمان ساده‌تر و قابل فهم‌تر می‌سازد.

مثالی از PCA - ادامه

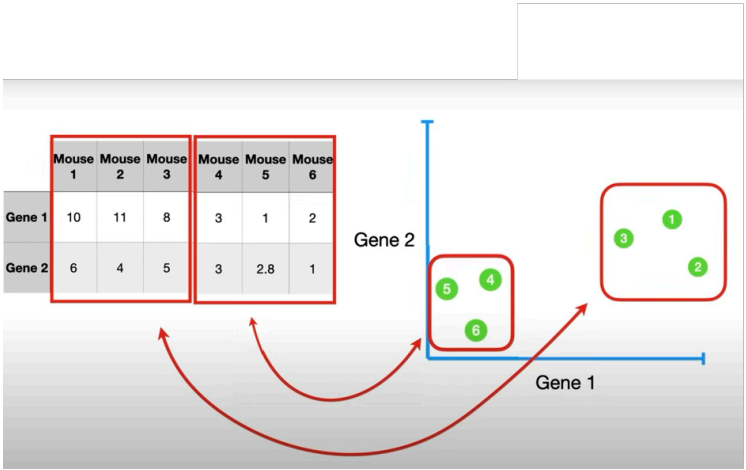
در ساده‌ترین حالت ممکن، می‌توان تنها یکی از ویژگی‌ها (ژن ۱) را نمایش داد که این حالت **تک بعدی** داده‌هاست.

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	1	2



- از همین نمایش ساده نیز می‌توان متوجه شد که موش‌های شماره ۲،۱ و ۳ نسبت به موش‌های شمار ۵،۴ و ۶ شباهت بیشتری با یکدیگر دارند.

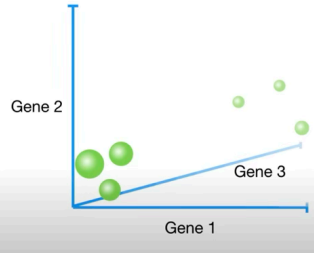
مثالی از PCA – ادامه



مثالی از PCA – ادامه

- با اضافه شدن یک ویژگی دیگر (ژن ۳)، داده‌های ما به **فضای سه‌بعدی** منتقل می‌شوند.

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2



مثالی از PCA - ادامه

- حالا هر نقطه روی این نمودار سه‌بُعدی، موقعیت یکی از موش‌ها را بر اساس مقادیر هر سه ژن (ژن ۱، ژن ۲، و ژن ۳) نشان می‌دهد.
- فکر کردن و کار کردن با داده‌ها در ۳ بُعد ممکن است، اما تصور کنید اگر ۱۰۰ یا ۱۰۰۰ ژن داشته باشیم! کشیدن نمودار و درک مستقیم داده‌ها غیرممکن می‌شود. PCA دقیقاً برای **ساده‌سازی و کار با داده‌ها در این ابعاد بالا** طراحی شده است.

محاسبه‌ی PCA

برای اینکه متوجه بشویم PCA چیست و چطور کار می‌کند، اجازه بدهید که برگردیم به همان جدول ابتدایی که شامل تنها دو ژن بود.

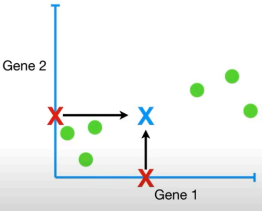
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

محاسبه‌ی PCA – ادامه

این اسلاید، اولین گام عملی در PCA را نشان می‌دهد. در این مرحله:

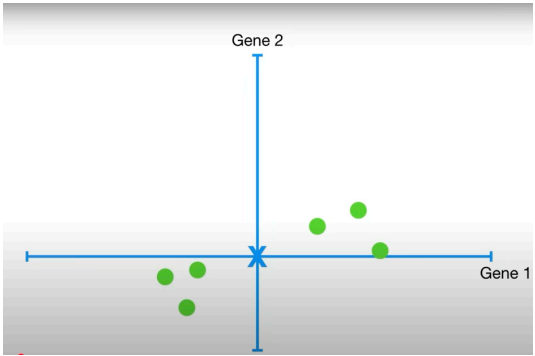
- ما میانگین مقادیر هر ژن (هر ویژگی) را به طور جداگانه حساب می‌کنیم. (میانگین ژن ۱ و میانگین ژن ۲).
- علامت‌های ضربدر قرمز روی محورها، این میانگین‌ها را نشان می‌دهند.
- نقطه ضربدر آبی رنگ در وسط نمودار، همان "مرکز" مجموعه داده‌های ماست که با استفاده از این میانگین‌ها مشخص می‌شود.

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1



محاسبه‌ی PCA – ادامه

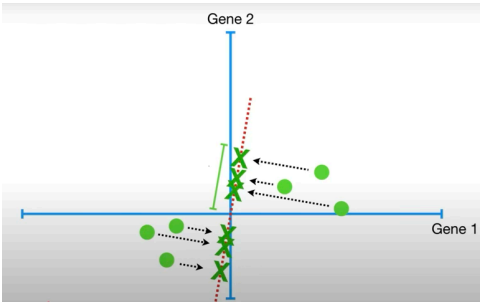
- بعد از پیدا کردن این مرکز، داده‌ها را طوری منتقل می‌کنیم که مرکزشان به مبدأ مختصات برود.



محاسبه‌ی PCA - ادامه

بعد از اینکه داده‌ها را در مرکز قرار دادیم (مرحله قبل)، حالا PCA شروع می‌کند به بررسی جهت‌های مختلف.

- در این تصویر، PCA یک خط فرضی (خط چین قرمز) را در نظر گرفته است.
- نقاط داده اصلی (نقاط سبز) روی این خط فرضی "تصویر" شده‌اند. "تصویر کردن" یعنی انگار از هر نقطه به صورت عمود به آن خط، خطی می‌کشیم و می‌بینیم کجا به خط فرضی می‌رسد. نقاط ضربدر سبز، محل تصویر شده نقاط اصلی روی این خط هستند.

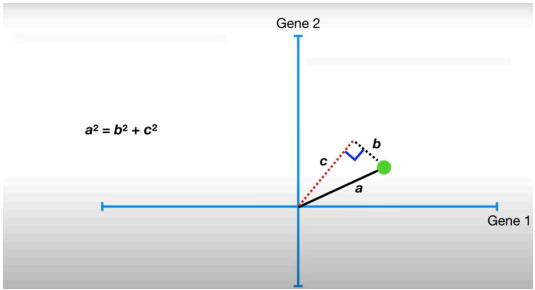


محاسبه‌ی PCA – ادامه

- PCA دنبال خطی می‌گردد که وقتی داده‌ها روی آن تصویر می‌شوند، نقاط تصویر شده بیشترین فاصله و پخش شدگی را از مرکز داشته باشند. (خط سبز کنار محور عمودی، میزان این پخش شدگی روی خط چین قرمز را نشان می‌دهد).
- به زبان ساده، PCA مثل این است که سعی کند بهترین زاویه را برای عکس گرفتن (مانند مثال قبلی) پیدا کند؛ زاویه‌ای که وقتی از آن نگاه می‌کنیم، داده‌ها (نقاط) تا جای ممکن از هم فاصله داشته باشند و روی هم نیفتند (بیشترین پراکندگی را داشته باشند). PCA این کار را با امتحان کردن خطوط مختلف و محاسبه میزان پخش شدگی نقاط تصویر شده روی آن‌ها انجام می‌دهد.

پیدا کردن بهترین خط (باکمک قضیه فیثاغورس)

برای اینکه بهتر بفهمیم PCA چطور آن **بهترین خط** را برای نمایش داده‌ها پیدا می‌کند، بیایید نگاهی به این اسلاید و نکته ریاضی ساده‌ای که نشان می‌دهد بیندازیم.



پیدا کردن بهترین خط (باکمک قضیه فیثاغورس)

یک نقطه داده (سبز) را در فضای دوبُعدی (بعد از مرحله مرکز قرار دادن) و یک خط فرضی که می‌خواهیم داده‌ها را روی آن "تصویر" کنیم (خط چین قرمز) می‌بینید.
حالا به اندازه‌های روی شکل توجه کنید:

- a: فاصله خود نقطه داده اصلی (نقطه سبز) از مرکز است. این فاصله برای هر نقطه، یک مقدار ثابت است.
- b: فاصله‌ی عمودی نقطه اصلی تا خط فرضی است. در واقع، این اندازه "خطای" تصویر کردن نقطه روی این خط را نشان می‌دهد؛ یعنی چقدر با تصویر کردن، نقطه اصلی را "گم" کرده‌ایم.
- c: فاصله نقطه‌ای که روی خط فرضی "تصویر" شده است، از مرکز است. این نشان‌دهنده میزان "پراکندگی" یا "اطلاعاتی" است که نقطه روی این خط دارد.

پیدا کردن بهترین خط (باکمک قضیه فیثاغورس)

این سه اندازه (a، b، c) یک مثلث قائم‌الزاویه تشکیل می‌دهند. بر اساس قضیه معروف فیثاغورس، رابطه زیر همیشه بین آن‌ها برقرار است:

$$a^2 = b^2 + c^2$$

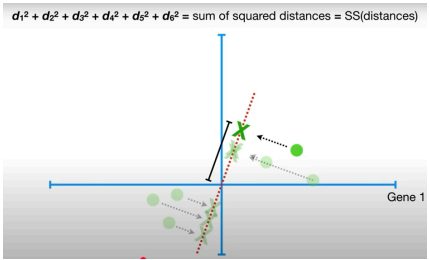
نکته کلیدی و بسیار مهم اینجاست:

از آنجایی که برای هر نقطه داده، a (فاصله اصلی از مرکز) یک عدد ثابت است، این رابطه به ما می‌گوید که:

- اگر c (یعنی میزان پراکندگی نقطه روی خط فرضی) **بیشتر** شود، آنگاه b (یعنی میزان خطا یا فاصله نقطه از آن خط) حتماً باید **کمتر** شود.
- و کاملاً برعکس، اگر c **کمتر** شود، b **بیشتر** می‌شود.

جمع مربع فاصله‌ها

در اسلاید قبل دیدیم که برای یک نقطه، c فاصله نقطه تصویر شده روی خط از مرکز است و هرچه c بیشتر باشد، b (خطا) کمتر است.



PCA این ایده را برای تمام نقاط داده اعمال می‌کند. فرمول بالا در واقع دارد **مربع فاصله هر یک از ۶ نقطه تصویر شده (ضربدرهای سبز) را از مرکز، جمع می‌کند.** (در اینجا d_i همان c برای موش iام است.)

$$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2 = SS(distances)$$

جمع مربع فاصله‌ها

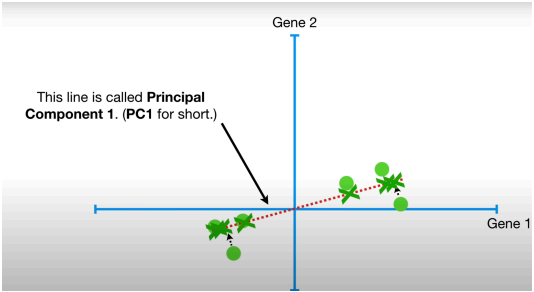
این **جمع مربع فاصله‌ها** (Sum of Squared Distances) یک عدد است که نشان می‌دهد مجموعاً چقدر داده‌ها روی این خط فرضی خاص پخش شده‌اند یا چقدر اطلاعات (تنوع) کلی در این جهت وجود دارد. خط سیاه با دو فلش کنار محور عمودی نیز به صورت تصویری، همین میزان پراکندگی کلی را روی خط چین قرمز نشان می‌دهد.

◀ کاری که PCA انجام می‌دهد این است:

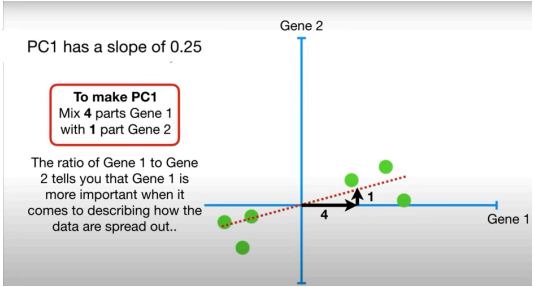
این **جمع مربع فاصله‌ها (میزان پراکندگی کلی)** را برای تمام خطوط ممکنی که از مرکز داده‌ها می‌گذرند محاسبه می‌کند.

بهترین خط

خطی که در آن این عدد (جمع مربع فاصله‌ها یا همان پراکندگی کلی) حداکثر مقدار ممکن را داشته باشد، همان **بهترین خط** ماست! این خط به عنوان **اولین مولفه اصلی (Principal Component۱)** انتخاب می‌شود. این خط، جهت‌گیری‌ای است که بیشترین اطلاعات و تنوع داده‌ها را در خود نگه داشته و اگر داده‌ها را روی آن تصویر کنیم، بیشترین پراکندگی را خواهند داشت و در عین حال کمترین خطا را نسبت به نقاط اصلی دارند.



شیب خط



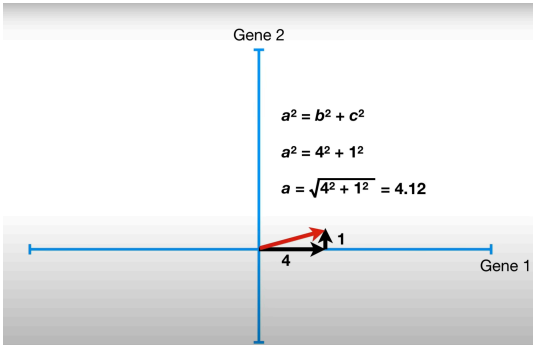
با به دست آوردن شیب خط اولین مولفه اصلی (PC۱) ، در واقع داریم می‌فهمیم که نسبت تغییر در جهت ژن ۲ نسبت به ژن ۱ در راستای این خط کلیدی چقدر است.

شیب خط

همانطور که در تصویر اسلاید قبل می‌بینید، شیب خط PC۱ برابر با 0.25 است. این عدد به ما می‌گوید:

- برای اینکه روی این خط PC۱ حرکت کنیم، به ازای هر ۴ واحدی که در راستای محور افقی (ژن ۱) جلو می‌رویم، باید ۱ واحد در راستای محور عمودی (ژن ۲) بالا برویم تا روی خط بمانیم. (تصویر فلش با عدد ۴ و ۱ این نسبت حرکت را نشان می‌دهد).
- شیب خط، که نسبت تغییر عمودی به تغییر افقی است، دقیقاً برابر با همین نسبت است.
- شیب خط PC۱ (عدد 0.25) بیان عددی دقیقی از **جهت این مولفه اصلی** و همان **نسبت مخلوط یا وزن‌های ژن‌ها** در ساخت PC۱ است که قبلاً درباره آن صحبت کردیم. این عدد نشان می‌دهد که خط PC۱ چقدر به سمت محور ژن ۱ متمایل است و چقدر به سمت محور ژن ۲.

محاسبه طول برداری



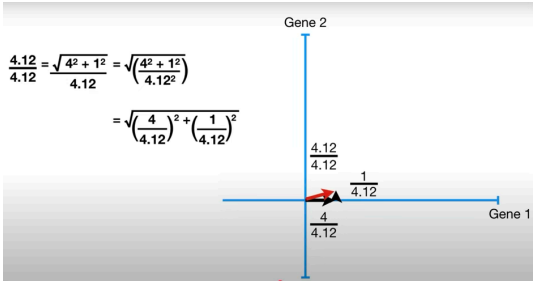
مرحله بعدی محاسبه طول برداری است که جهت PC۱ را نشان می‌دهد. فلش سیاه رنگ، یک **بردار** است که از مرکز شروع شده و به سمت جهت PC۱ می‌رود. این بردار با حرکت ۴ واحد در راستای ژن ۱ و ۱ واحد در راستای ژن ۲ ساخته شده است.

محاسبه طول برداری

محاسباتی که در تصویر اسلاید قبل می‌بینید، با استفاده از همان قضیه فیثاغورس که قبلاً دیدیم، طول این بردار را حساب می‌کنند.
فلش قرمز رنگی که در همان جهت قرار دارد، **بردار واحد (Unit Vector)** نامیده می‌شود. بردار واحد، برداری است که همان جهت را نشان می‌دهد اما طولش دقیقاً ۱ است. در PCA معمولاً جهت مولفه‌های اصلی را با بردارهای واحد نشان می‌دهند.
اهمیت این مرحله در این است که مولفه‌های همین بردار واحد (فلش قرمز) هستند که به عنوان بارهای مولفه اصلی (Principal Component Loadings) شناخته می‌شوند. این بارها نسخه‌ی استاندارد شده همان نسبت‌های ۴ و ۱ هستند و به ما می‌گویند هر ژن چقدر **سهم** یا **بار** در ساخت این مولفه اصلی (PC۱) دارد.

استاندارد کردن بردار جهت

گام بعدی **استاندارد کردن بردار جهت** است تا طول آن دقیقاً برابر با ۱ شود (همان **بردار واحد** که با فلش قرمز نشان داده شده). برای این کار، باید مولفه‌های بردار اصلی (یعنی ۴ و ۱) را بر طول خودش (۴.۱۲) تقسیم کنیم.



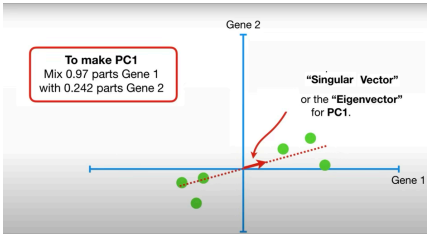
استاندارد کردن بردار جهت

- محاسباتی که در بالای تصویر اسلاید قبل می‌بینید، همین کار را نشان می‌دهند و تأیید می‌کنند که با تقسیم کردن بردار اصلی بر طولش (۴.۱۲)، به یک بردار با طول ۱ می‌رسیم.
- اعدادی که کنار محورها می‌بینید ($\frac{4}{4.12}$ روی محور ژن ۱ و $\frac{1}{4.12}$ روی محور ژن ۲)، در واقع مولفه‌های همین بردار واحد PC۱ هستند.

به این مولفه‌های بردار واحد PC1، **بارهای مولفه اصلی** (Principal Component Loadings) گفته می‌شود. این بارها همان وزن‌های نهایی و استاندارد شده ژن ۱ و ژن ۲ در ساختار مولفه اصلی اول (PC۱) هستند.

این بارها به ما می‌گویند که هر ژن چقدر **سهم** یا **اهمیت** در ساختار PC۱ دارد. همانطور که می‌بینید، بار ژن ۱ خیلی بزرگتر از بار ژن ۲ است (۰/۹۷ در مقابل ۰/۲۴)، که تأیید می‌کند ژن ۱ نقش بسیار مهم‌تری در **جهت اصلی پراکندگی داده‌ها** داشته است.

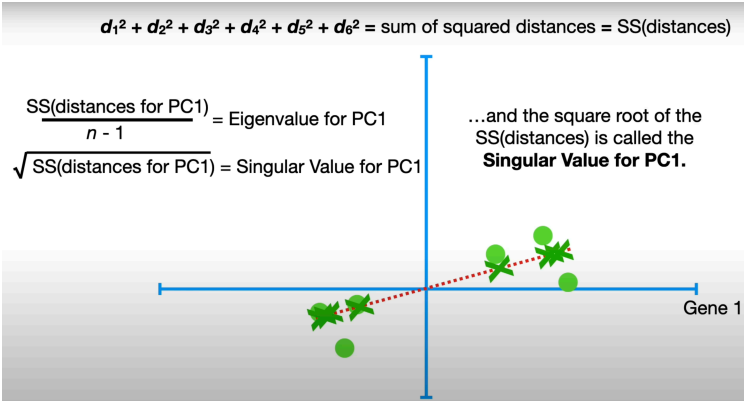
بردار ویژه



در اسلاید قبل دیدیم که با استاندارد کردن بردار جهت PC1، به اعدادی رسیدیم (حدوداً ۰/۹۷ برای ژن ۱ و ۰/۲۴ برای ژن ۲) که به آن‌ها بارهای مولفه اصلی می‌گوییم. این اسلاید نشان می‌دهد که این بارها دقیقاً همان اعدادی هستند که در **دستورالعمل دقیق‌تر مخلوط** برای ساختن PC۱ استفاده می‌شوند:

۰/۹۷ واحد از ژن ۱ را با ۰/۲۴ واحد از ژن ۲ مخلوط کن تا PC۱ را بسازی!

مقدار ویژه


$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2$ = sum of squared distances = SS(distances)
$\frac{\text{SS(distances for PC1)}}{n - 1}$ = Eigenvalue for PC1
$\sqrt{\text{SS(distances for PC1)}}$ = Singular Value for PC1
...and the square root of the SS(distances) is called the **Singular Value for PC1.**
Gene 1

مقدار ویژه

حالا که جهت PC۱ و بارهای آن را می‌دانیم، اسلاید بالا، نتیجه تصویر کردن داده‌های اصلی (نقاط سبز) روی خط PC۱ را به ما نشان می‌دهد. ضربدرهای سبز روی خط چین قرمز، همان نقاط تصویر شده هستند.
فرمول بالای اسلاید قبل ((SS(distances for PC۱)، همان جمع مربع فاصله‌های نقاط تصویر شده (ضربدرهای سبز) از مرکز، روی خط PC۱ است که قبلاً هم دیدیم. این عدد نشان‌دهنده میزان کل پراکندگی یا واریانس داده‌ها فقط در جهت PC۱ است.
در PCA، به این میزان پراکندگی (وقتی بر تعداد نمونه‌ها منهای یک تقسیم شود)، یک نام خیلی مهم می‌دهند:
مقدار ویژه (Eigenvalue)

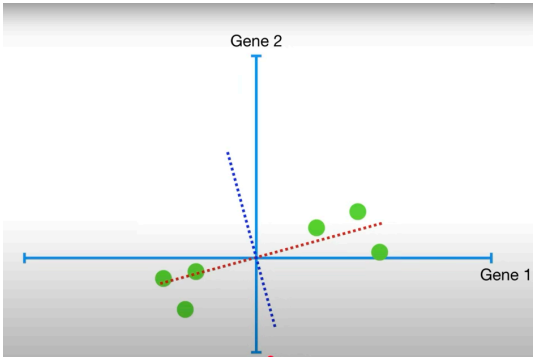
مقدار ویژه

چرا این مقدار ویژه (Eigenvalue) مهم است؟

- چون مقدار ویژه هر مولفه اصلی به ما می‌گوید که آن مولفه به تنهایی، **چقدر از کل اطلاعات و تنوع (واریانس کلی)** موجود در تمام داده‌های اصلی را در خود نگه داشته و توضیح می‌دهد.
- مقدار ویژه بزرگتر برای PC۱ یعنی PC۱ توانسته بخش بسیار بزرگی از واریانس کلی داده‌ها را به خود اختصاص دهد و در نتیجه یک مولفه خیلی مهم برای خلاصه کردن داده‌هاست. با مقایسه مقادیر ویژه مولفه‌های مختلف (PC۱، PC۲ و ...)، می‌توانیم بفهمیم کدام مولفه‌ها بیشترین اطلاعات را دارند و کدام را می‌توانیم کنار بگذاریم.

دومین مولفه اصلی

حالا که PC۱ را کامل متوجه شدیم، بیایید روی PC۲ کار کنیم!!!

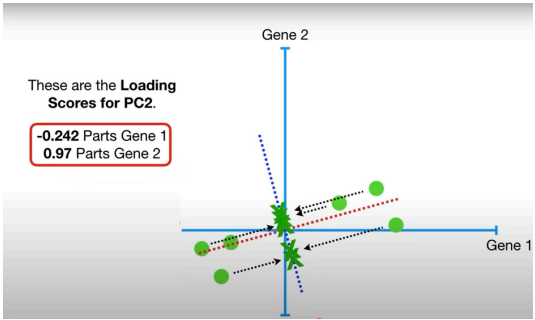


دومین مولفه اصلی

بعد از اینکه اولین مولفه اصلی (PC۱) را پیدا کردیم که مهم‌ترین جهت پراکندگی داده‌ها و بیشترین واریانس را به خود اختصاص داده بود، PCA به سراغ پیدا کردن مولفه‌های اصلی بعدی می‌رود تا باقی‌مانده اطلاعات و تنوع در داده‌ها را پوشش دهد.

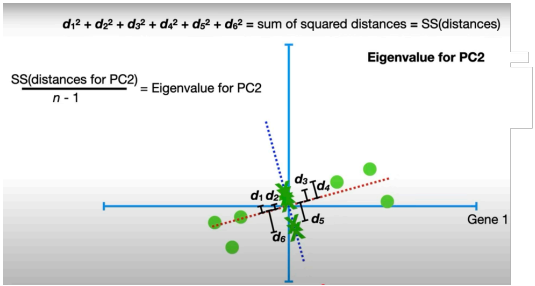
- در این مرحله، هدف پیدا کردن دومین مولفه اصلی (Principal Component ۲ یا PC۲) است.
- ویژگی بسیار مهم و کلیدی PC۲ در یک فضای دوبُعدی مانند مثال ساده ما (ژن ۱ و ژن ۲) این است که PC۲ **همیشه باید عمود (Perpendicular) بر اولین مولفه اصلی (PC۱)** باشد!
- PC۲ جهتی را پیدا می‌کند که بیشترین **باقی‌مانده پراکندگی (واریانس باقی‌مانده)** در داده‌ها را در خود جای داده است؛ یعنی آن بخشی از تنوع داده‌ها که توسط PC۱ توضیح داده نشده است.

دومین مولفه اصلی



- PC۲ هم بردار ویژه و بارهای مخصوص به خودش را دارد که نشان می‌دهند در این جهت، ژن‌های اصلی با چه وزن‌هایی ترکیب شده‌اند.

دومین مولفه اصلی



- میزان واریانسی که PC۲ پوشش می‌دهد، با "مقدار ویژه" مربوط به PC۲ اندازه‌گیری می‌شود. مقدار ویژه PC۲ از مقدار ویژه PC۱ کوچکتر است.

پیدا کردن مولفه‌های اصلی بیشتر

- اگر داده‌های اولیه ما ابعاد بیشتری داشته باشند (مثلاً ۳ متغیر یا بیشتر)، PCA همین فرآیند را تکرار می‌کند. به ترتیب مولفه اصلی سوم ()، چهارم (PC۴) و همینطور تا تعداد متغیرهای اولیه پیدا می‌شوند.
- هر مولفه جدید (PC_k) جهتی را پیدا می‌کند که بیشترین باقی‌مانده واریانس را نسبت به تمام مولفه‌های قبلی خود (PC_1 تا PC_{k-1}) داشته باشد.
- هر PC_k عمود بر تمام مولفه‌های اصلی قبلی خود است.
- هر PC_k هم، بردار ویژه، بارها (Loadings) و مقدار ویژه مخصوص به خودش را دارد. مقادیر ویژه به ترتیب برای مولفه‌های بعدی کوچکتر و کوچکتر می‌شوند، چون هر بار داریم از باقی‌مانده اطلاعات می‌گیریم.

نتیجه نهایی

- در پایان، PCA مجموعه‌ای از مولفه‌های اصلی (به تعداد متغیرهای اولیه) به ما می‌دهد: PC_1، PC_2، PC_3، و غیره.
- این مولفه‌ها **بر هم عمودند**.
- بر اساس میزان **واریانسی** که پوشش می‌دهند (از طریق **مقادیر ویژه شان**) به ترتیب مرتب شده‌اند: PC_1 مهم‌ترین (بیشترین واریانس)، PC_2 بعدی، و همینطور الی آخر.
- هر کدام از این PC ها، یک **جهت** یا **بعد جدید** در داده‌ها هستند که ترکیبی از متغیرهای اصلی با بارهای مشخص خود هستند.

نتیجه نهایی - ادامه