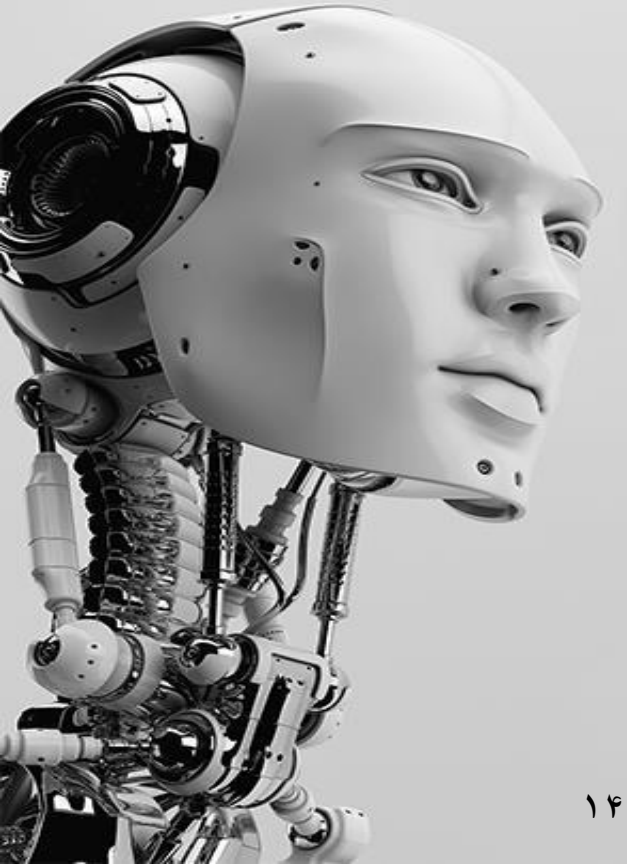
بخش اول » ریاضیات پایه و جبر خطی برای یادگیری‌ماشین

03

بخش اول » ریاضیات پایه و جبر خطی برای یادگیری‌ماشین

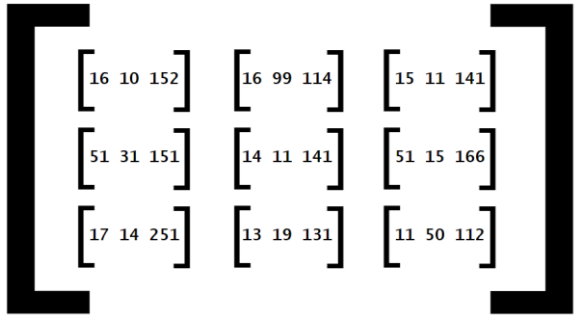
03

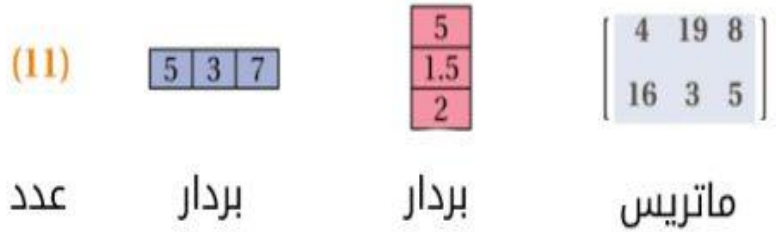


سید ایمان غفوریان - عضو هیات علمی دانشگاه آزاد مشهد-زمستان ۱۴۰۲

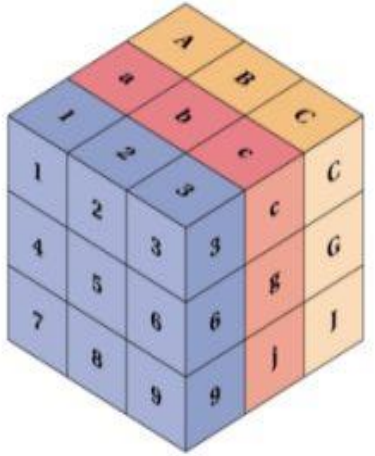
Tensorتنسور ، Matrixماتریس ، Vectorsبردار ، عدد Scalar

پایه‌ی عملیاتِ یادگیری‌ماشین و یادگیری‌عمیق ، **اعداد** هستند. این اعداد با قرار گرفتن در کنارِ هم ، **بردارها** را می‌سازند و بردارها در کنار یکدیگر قرار می‌گیرند و **ماتریس‌ها** را تشکیل می‌دهند. تصویر زیر یک نمونه **تنسور** است که سه بُعد دارد. به بیانی دیگر تصویرِ بالا یک ماتریس است که هر کدام از خانه‌های آن، خود یک بردار هستند. در واقع تنسور یک بُعد بیشتر از ماتریس داردعدد ۰ بُعد دارد، بردار ۱ بُعدی است، ماتریس ۲ بُعدی و در نهایت تنسور ۳بُعد دارد.





عدد بردار بردار ماتریس

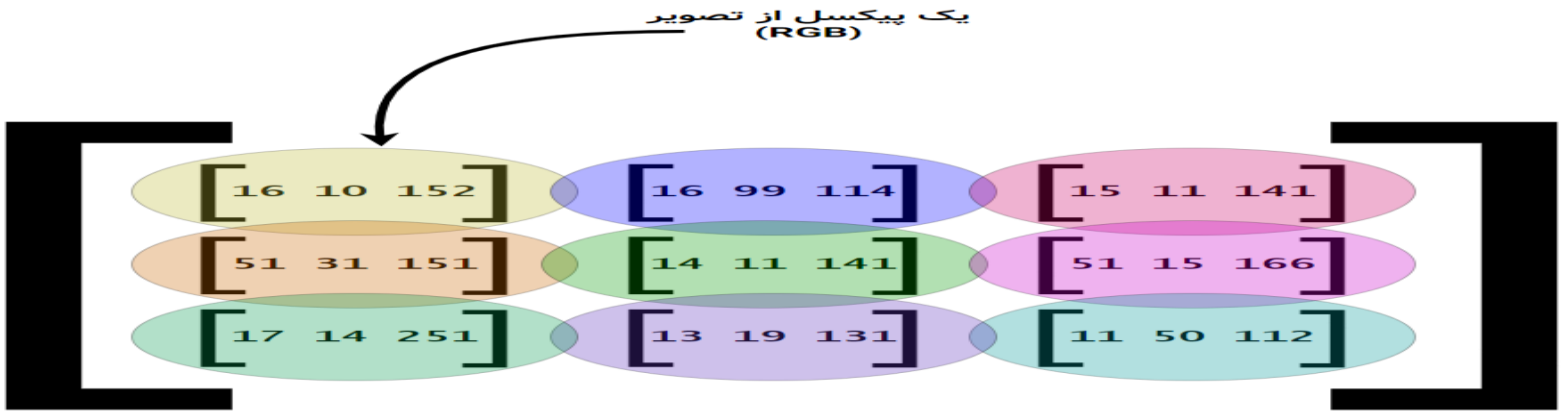


تنسور

Tensorتنسور ، Matrixماتریس ، Vectorsبردار ، عدد Scalar

در واقع یک ماتریس از مشتریان ساخته‌ایم که هر مشتری یک بردار است و این یکی از کاربردهای اساسی بردار و ماتریس در عملیات یادگیری ماشین و داده‌کاوی است. اما در مورد کاربردِ تنسور tensor می‌توان مثالِ یک عکس (مانند JPEG)را آورد. فرض کنید می‌خواهیم یک تصویر را به کامپیوتربه وسیله‌ی تنسور وارد کنیم. در واقع بایستی کاری کنیم که این تصویر برای کامپیوترو به تبعِ آن الگوریتم‌های داده‌کاوی قابلِ فهم باشد. عرض و ارتفاع تصویر که مانند یک ماتریس است و هر خانه ماتریس هم یک پیکسل (pixel)از آن تصویر است. همان‌طور که می‌دانید در سیستم RGB هر پیکسل از مجموعه رنگ قرمز، سبز و آبی (RGB)تشکیل شده است. پس هر کدام از خانه‌های این ماتریس را می‌توان ترکیبی از این سه رنگ دانست.

ما یک تصویر را به یک تنسورtensor تبدیل کرده‌ایم. هر کدام از خانه‌های این ماتریس، نمایان‌گرِ یک پیکسل است، که خود از سه رنگ (یک بردارِ سه عضوی) تشکیل شده است.



ماتریس‌ها و کاربرد آن‌ها در داده‌کاوی و یادگیری ماشین

فرض کنید در یک آتش‌نشانی در منطقه‌ی جنگلی مشغول به کار هستید. در این منطقه هر روز احتمال آتش‌سوزی کوچک وجود دارد که در صورت عدم رسیدگی می‌تواند به آتش‌سوزی بزرگ در جنگل تبدیل شود. آتش‌نشان‌ها به صورت تجربی می‌دانند که قبل از وقوع آتش‌سوزی ممکن است تغییراتی در آب و هوا رخ دهد. برای مثال می‌دانند که اگر در روز قبل دمای هوا بالا باشد و در چند هفته‌ی گذشته بارانی نباریده باشد و همچنین باد شروع به وزیدن کرده باشد، این احتمال هست که فردا آتش‌سوزی داشته باشیم.

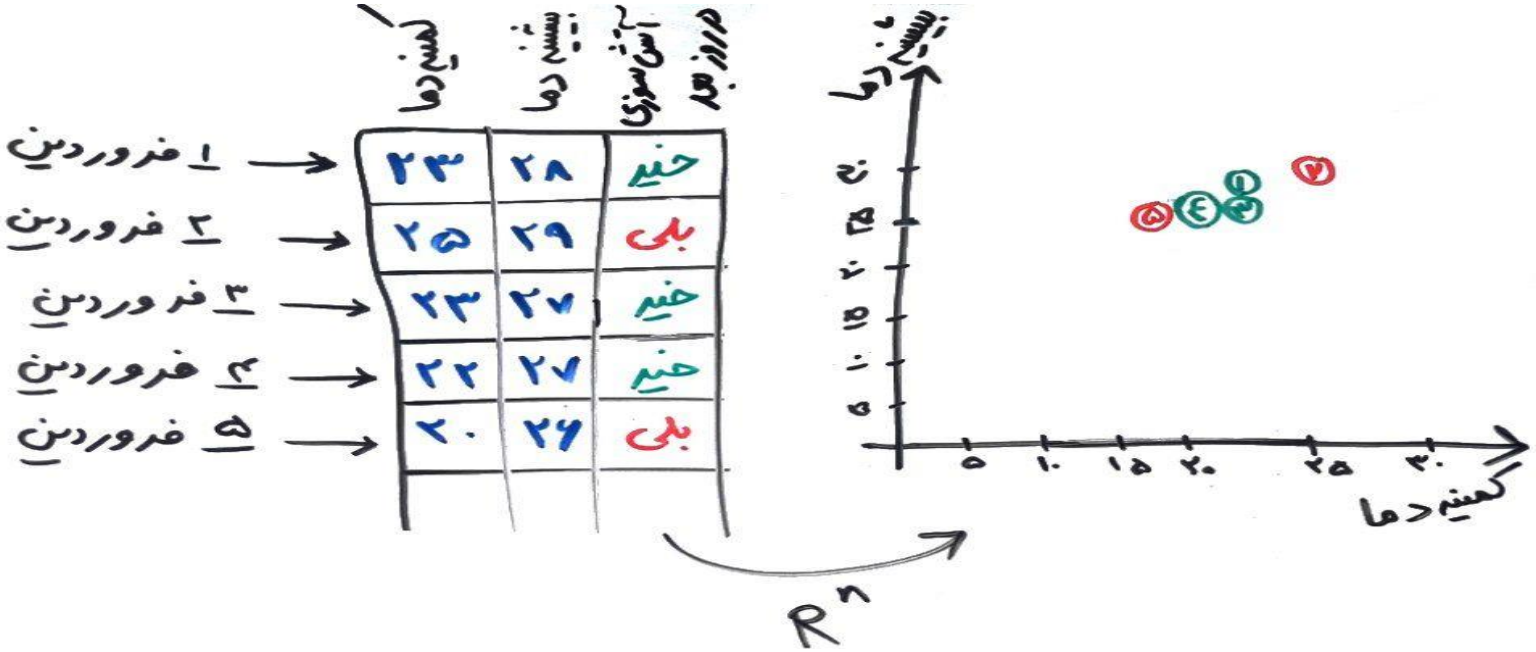
مثال زیر می‌تواند با استفاده از داده‌ها علمی‌تر و دقیق‌تر باشد. یک متخصص علم داده می‌تواند با **جمع‌آوری داده‌ها، ماتریسی** به شکل زیر بسازد، در مثال زیر هر کدام از **سطرها**، **یک روز** را نشان می‌دهد. هر **روز** ۷ **ویژگی** دارد، به همراه یک **ستون آخر** که نشان می‌دهد فردای آن روز، در جنگل آتش‌سوزی رخ داده است یا خیر در آینده خواهیم دید که این ماتریس می‌تواند به الگوریتم‌های یادگیری ماشین تزریق شود و این الگوریتم‌ها از روی داده‌ها یادگیری را انجام دهند. سپس برای روزهای آینده، الگوریتم، با مشاهده‌ی ویژگی‌های آن روز، پیش‌بینی کند که آیا فردا آتش‌سوزی رخ خواهد داد یا خیر.

۱ فردا →	۲۳	۲۸	۲۵	۱	۰	۸	۱۷	خیر
۲ فردا →	۲۵	۲۹	۲۶	۰	۰	۷	۱۷	بلی
۳ فردا →	۲۳	۲۷	۲۶	۰	۰	۶	۱۷	خیر
۴ فردا →	۲۲	۲۷	۲۵	۱	۰	۶	۱۸	خیر
۵ فردا →	۲۰	۲۶	۲۱	۰	۰	۳	۱۶	بلی
	⋮							

ماتریس‌ها و کاربرد آن‌ها در داده‌کاوی و یادگیری ماشین

همان‌طور که می‌بینید این ماتریس، مجموعه‌ای از داده‌ها را به صورت ساختاریافته در خود ذخیره کرده است. هر **سطر** از این ماتریس یک نمونه **instance**نامیده می‌شود که نشان‌دهنده‌ی یک **روز** است. همچنین هر **ستون** از این ماتریس یک **بُعد Dimension**یا **ویژگی Feature**نامیده می‌شود، زیرا هر ستون یک ویژگی از یک نمونه را مشخص می‌کند.

ماتریس می‌تواند در یک فضای اقلیدسی نمایش داده شود. البته ما در **صفحه** نمی‌توانیم ابعاد بالاتر از **سه بُعد** را رسم کنیم. پس برای **سادگی** فرض کنید که ماتریس بالا، به جای ۷ **بُعد**، فقط ۲ **بُعد** دارد. اگر بخواهیم این شکل را بر روی یک فضای اقلیدسی نمایش دهیم، هر **سطر** (نمونه) یک نقطه در یک فضای ۲ بُعدی (مثلاً x1و x2) می‌شود. چیزی شبیه به شکل زیر:



همان‌طور که مشاهده می‌کنید، هر سطر به یک نقطه در فضا نگاشت شده است. این فضا که همان فضای **دکارتی** است، یکی از اصول جبرخطی بوده و کاربرد بسیار زیادی از یادگیری ماشین و داده‌کاوی دارد.

Normنرم بردار یا ماتریس چیست؟

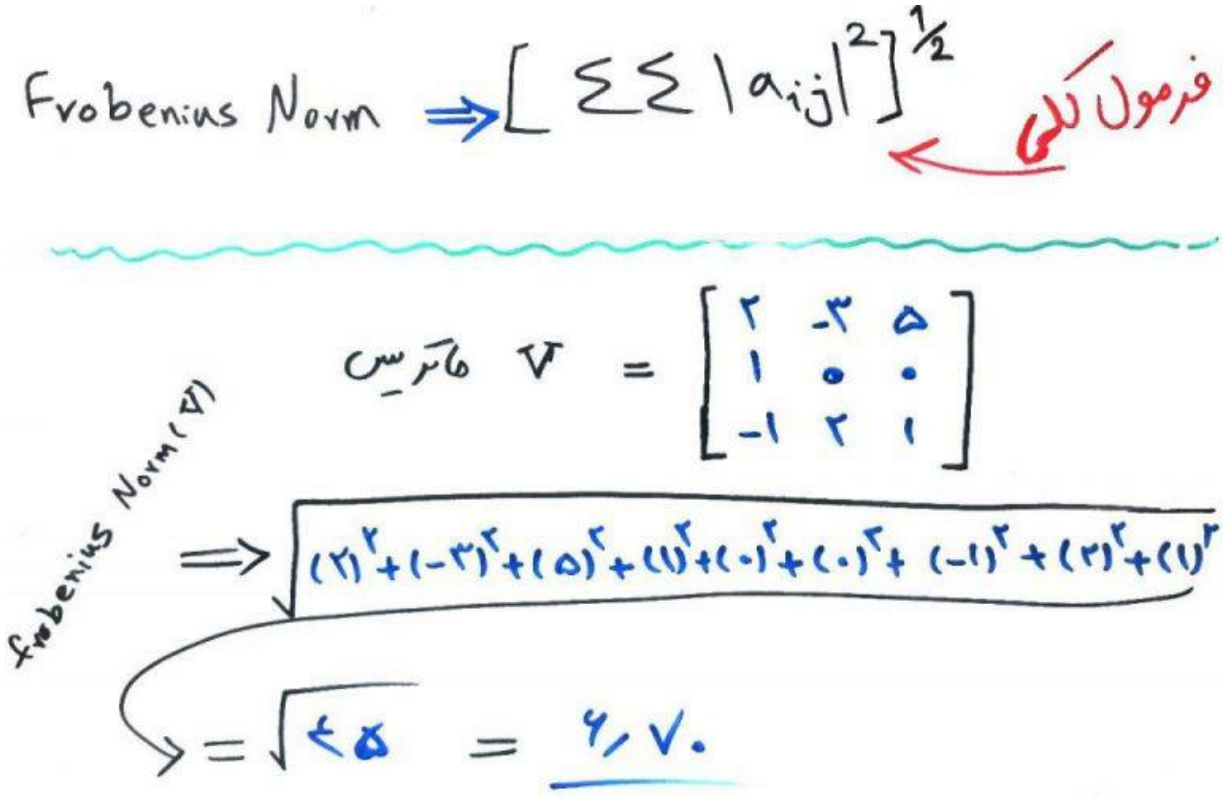
وقتی بخواهند اندازه‌ی یک بردار vectorیا ماتریس matrix را محاسبه کنند از عبارتِ نُرم یا همان normاستفاده می‌کنند. نُرمِ یک بردار یا ماتریس، کاربردِ فراوانی مخصوصاً در یادگیری ماشین و شبکه‌های عصبی عمیق deep neural networkدارد.

نحوه‌ی محاسبه‌ی نرم normمتفاوت است، به همین دلیل normها اسامی مختلفی دارند. مثلا نُرمِ اُقلیدوسی یا L1 Normو یا L2 Normکه ممکن است در بسیاری از مراجع داده‌کاوی و یادگیری ماشین دیده باشید، از پرکاربردترین نُرم‌ها هستند. برای آشنایی با آن‌ها به یکی از معروف‌ترین نُرم‌ها که همان نُرمِ اُقلیدوسی یا L2 normاست می‌پردازیم. شکل زیر را نگاه کنید:

همان‌طور که در شکل می‌بینید ما یک بردار vectorداریم و هر کدام از عناصرِ آن را به توان ۲ رساندیم و با هم جمع کردیم، سپس از مجموعِ آن‌ها رادیکال گرفتیم و در نهایت به یک عدد (۶/۲۴۵) رسیدیم. این عدد در واقع نشان دهنده‌ی بزرگیِ این بردار است. پس در واقع می‌توان گفت نُرم، یک نوع تخمین برای به دست آوردن بزرگی بردار است.

Normنرم بردار یا ماتریس چیست؟

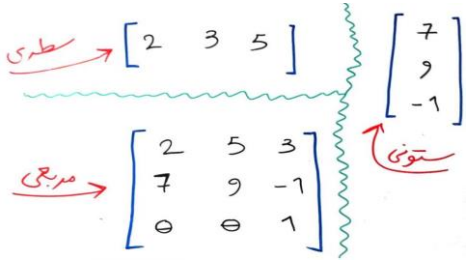
می‌توان مقدارِ نُرم را برای ماتریس هم محاسبه کرد. یکی از معروف‌ترین نُرم‌ها برای ماتریس، نُرم فروبنیوس frobenius norm است که چیزی شبیه به L2 normبرای بردارهاست. به شکل زیر نگاه کنید:



فرمول ساده است، تمامی اعداد داخل ماتریس (هر سطر به ازای هر ستون) را به توان ۲ می‌رسانیم و با یکدیگر جمع می‌کنیم. سپس از آن‌ها رادیکال می‌گیریم.

در واقع اینجا هم یک ماتریس را به یک عدد تبدیل کردیم. با این کار به صورت تخمینی می‌توانیم متوجه شویم که اندازه‌ی یک ماتریس چقدر بزرگ است و یا اینکه ماتریس‌ها را با هم مقایسه کنیم.

انواع ماتریس و ویژگی‌های مختلف آن‌ها



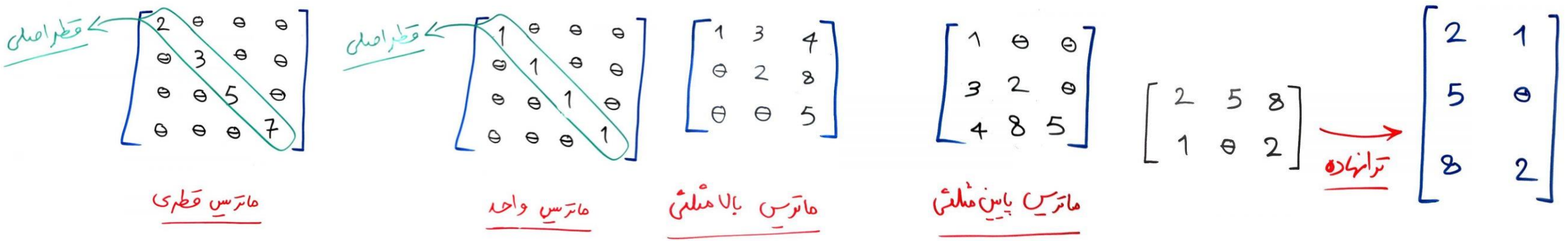
شکل زیر سه نوع ماتریس هستند که به ترتیب ماتریس‌های سطری، ستونی و ماتریس مربعی squared matrixنامیده می‌شوند:

۱–در شکل زیر یک ماتریس قطری diagonal matrixداریم، یعنی ماتریسی که همه عناصر آن به جز قطر اصلی صفر است.

۲–ماتریس واحد identity matrixهم یک ماتریس قطری است که اعداد موجود در قطر اصلی آن، ۱ هستند.

۳–شکل‌های زیر ماتریس‌های مثلثی triangle matrixنامیده می‌شوند. دو نوع ماتریس مثلثی وجود دارد. پایین مثلی و بالا مثلثی. ماتریس بالا مثلثی ماتریسی است که اعدادِ بالاییِ قطرِ اصلیِ آن موجود بوده و بقیه صفر هستند و ماتریس پایین مثلثی هم برعکس

۴–و در آخر ترانهاده‌ی یک ماتریس matrix transpose، ماتریسی است که جای سطر و ستون آن را با هم عوض می‌کنیم .



چرا ماتریس‌ها در علوم داده مهم هستند؟

ماتریس‌ها یکی از عناصر اصلی در ریاضیات و جبرخطی هستند. بدون استفاده از ماتریس‌ها، نمی‌توان ساختارهای داده‌ای پیچیده و ترکیبی را ساخت. در علوم کامپیوتر و محاسبات هم ماتریس‌ها جایگاه ویژه‌ای دارند. عملیات محاسباتی سنگین مانند پردازش‌های گرافیک (مانند بازی‌ها یا ساخت انیمیشن‌ها) همه و همه از خاصیت پردازش ماتریس‌ها استفاده می‌کنند. کارت‌های گرافیک جدید هم به گونه‌ای طراحی شده‌اند که می‌توانند عملیات مختلف بر روی ماتریس‌های حجیم را در کسری از ثانیه اجرا کنند. نقش ماتریس‌ها در علوم داده data scienceنیز غیر قابل انکار است. با استفاده از جبر رابطه‌ای relational algebraمی‌توان هر گونه رابطه‌ای را در دنیای واقعی به جداول و ماتریس‌ها تبدیل کرد. برای مثال یک شرکت تولید لوازم خانگی را در نظر بگیرید که می‌خواهد اطلاعات مشتریان خود را در جدول یا ماتریسی ذخیره کند. چیزی مانند شکل زیر:

شماره مشتری	اس خریداری شده	مجموع مبلغ خرید	سن	تعداد فرزندان	استفاده از گارانتی	سال اولین خرید	سال آخرین خرید
Customer #1	یخچال	۲۰۰۰۰۰۰۰	۴۰	۲	۲	۱۳۹۲	۱۳۹۸
Customer #2	یخچال، تلویزیون	۳۲۰۰۰۰۰۰	۲۹	۱	۱	۱۳۹۹	۱۳۹۹
Customer #3	ماشین لباسشویی	۱۲۰۰۰۰۰۰	۲۰	۰	۰	۱۳۹۹	۱۳۹۹
...							

اگر به ماتریس بالا نگاه کنید، در هر سطر یک مشتری را ذخیره کرده‌ایم. هر ستون نشان دهنده‌ی یک «ویژگی» است. برای مثال شخص شماره‌ی ۱ را در نظر بگیرید. این شخص ۴۰ ساله است و دو فرزند دارد. یک یخچال از شرکت به قیمت ۲۰ میلیون تومان خریده است. همچنین ۲ مرتبه از گارانتی استفاده کرده است. همان‌طور که مشاهده می‌کنید تمامی افراد (سطرها) همه‌ی این ویژگی‌ها (ستون‌ها) را دارند. در اینجا ما مشتریان خود را به یک ماتریس (جدول) تبدیل کرده‌ایم و می‌توانیم با کمک این ساختار، داده‌های خود را ذخیره کنیم. همچنین بسیاری از الگوریتم‌های داده‌کاوی data miningو یادگیری ماشین machine learningهم برای اجرا و یادگیری الگو، از ورودی‌های ماتریس استفاده می‌کنند.به طور کلی ماتریس عنصری است که برای سازماندهی به اعداد و مدل‌سازی مفاهیم پیچیده با استفاده از ریاضیات استفاده می‌شود. با استفاده از ماتریس‌ها، ما می‌توانیم تقریباً همه‌ی عناصر موجود در طبیعت را در کامپیوتر ذخیره کرده و آن‌ها را پردازش کنیم. نوع پیچیده‌تری از ماتریس نیز وجود دارد که به آن تنسور tensorمی‌گویند.

سید ایمان غفوریان - عضو هیات علمی دانشگاه آزاد مشهد-زمستان ۱۴۰۲

معیارهای فاصله(Distance Measures)در یادگیری ماشین

در این قسمت می‌خواهیم به برخی از مشهورترین معیارهای فاصله بین نمونه‌ها در یک ماتریس از داده‌ها بپردازیم. مجموعه‌ی داده‌ی زیر که نمایانگر کاربران در یک اپلیکیشن پخش موسیقی مانند Spotifyاست را در نظر بگیرید:

Client/Song	Song1	Song2	Song3	Song4	Song5
101	4	0	0	1	2
102	3	1	0	0	0
103	0	0	0	3	3
104	0	1	1	1	3
105	6	6	1	6	0
106	0	0	0	0	3

هر سطر از ماتریسِ شکل بالا یک کاربر را نمایش می‌دهد و هر ستون، نشان‌دهنده‌ی تعداد دفعاتی است که آن کاربر، ی ک موسیقی خاص را پخش کرده است. برای مثال کاربر شماره‌ی ۱۰۱، تعداد ۴ مرتبه موسیقی شماره ۱ را پخش کرده و اصلا موسیقی شماره‌ی ۲ و ۳ را پخش نکرده است. همین کاربر به ترتیب ۱ مرتبه موسیقی شماره‌ی ۴ و ۲ مرتبه موسیقی شماره‌ی ۵ را پخش کرده است و به همین ترتیب برای بقیه‌ی کاربران (سطرها) می‌توان مجموعه‌ی داده‌ی بالا را تفسیر کرد.حال فرض کنید می‌خواهیم فاصله‌ی بین دو نمونه (در اینجا دو کاربر) یعنی دو سطر را اندازه بگیریم. کاربرانی که به یکدیگر نزدیک‌تر باشند، احتمالا سلیقه‌ی موسیقیایی مشابهی به یکدیگر دارند و می‌توان با استفاده از این اطلاعات، موسیقی‌هایی که کاربران مختلف احتمالاً علاقه دارند را به آن‌ها پیشنهاد داد.

برای محاسبه‌ی فاصله‌ی بین دو نمونه (هر نمونه را می‌توان یک بردار در نظر گرفت) در یک ماتریس، معیارهای فاصله‌ی متفاوتی وجود دارد که در ادامه به برخی از مشهورترین آن‌ها خواهیم پرداخت.

فاصله‌ی اقلیدسی Euclidean Distance

فاصله‌ی اقلیدسی بین دو نمونه (دو بردار) به صورت زیر محاسبه می‌شود:

$$Distance(p, q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

تعداد ابعاد (ستون‌ها)

معیارهای فاصله(Distance Measures)در یادگیری ماشین

فاصله‌ی اقلیدسی Euclidean Distance

فاصله‌ی اقلیدسی بین دو نمونه (دو بردار) به صورت مقابل محاسبه می‌شود:

تعداد ابعاد (ستون‌ها)

$$Distance(p,\ q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

در شکل بالا مشخص است که هر کدام از ابعاد (ستون‌ها) را برای هر کدام از نمونه‌ها (سطرها) از یکدیگر تفریق کرده و به توان ۲ می‌رسانیم. سپس حاصل این عبارات را جمع کرده و یک رادیکال از آن می‌گیریم. برای مثال اگر بخواهیم فاصله‌ی کاربر ۱۰۱ و ۱۰۳ در مجموعه‌ی داده‌ی بالا را با توجه به معیار فاصله‌ی اقلیدسی محاسبه کنیم، به صورت زیر این محاسبه انجام می‌گیرد:هر چقدر فاصله‌ی اقلیدسی بین نمونه‌ها کمتر باشد، یعنی آن دو نمونه به یکدیگر نزدیک‌تر هستند. فاصله‌ی اقلیدسی، رابطه‌ی مستقیمی با فاصله‌ی L2 Normدارد.

$$\sqrt{(0-4)^2 + (0-0)^2 + (0-0)^2 + (3-1)^2 + (3-2)^2} = 4.58$$

Song1 Song2 Song3 Song4 Song5

فاصله‌ی منهتن Manhattan Distance

فاصله‌ی منتهن که به فاصله‌ی تاکسی Taxicabیا فاصله‌ی بلوک‌های شهری City Blockنیز معروف است شبیه به فاصله‌ی اقلیدسی عمل می‌کند با این تفاوت که در فرمول، توانِ ۲ و رادیکال نداشته و فقط قدر مطلقِ اختلاف‌ها را محاسبه می‌کند.برای مثال اگر بخواهیم فاصله‌ی کاربر ۱۰۱ و ۱۰۳ در مجموعه‌ی داده‌ی بالا را با توجه به معیار فاصله‌ی منهتن محاسبه کنیم، به صورت زیر این محاسبه انجام می‌شود،فاصله‌ی منهتن رابطه‌ی مستقیمی با فرمول L1 Normنیز می‌گویند.

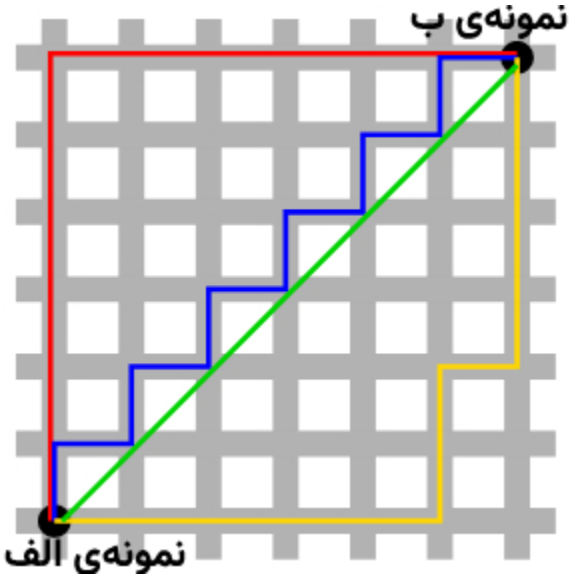
$$Distance(p,\ q) = \sum_{i=1}^{n} |q_i - p_i|$$

$$|0-4| + |0-0| + |0-0| + |3-1| + |3-2| = 7$$

Song1 Song2 Song3 Song4 Song5

معیارهای فاصله(Distance Measures)در یادگیری ماشین

به صورت خلاصه در شکل زیر تفاوت بین فاصله‌ی منهتن و اقلیدسی را برای حالتی که دو نمونه (دو سطر) و دو بُعد (دو ستون) داریم مشاهده می‌کنید:



در شکل بالا، خط مستقیم بین دو نمونه‌ی الف و ب (خط سبز رنگ)، فاصله‌ی اقلیدسی را نشان می‌دهد و خطوط دیگر (آبی، زرد و قرمز) هر کدام به نوعی نشان‌دهنده‌ی فاصله‌ی منهتن هستند.

فاصله‌ی مینکوفسکی Minkowski Distance

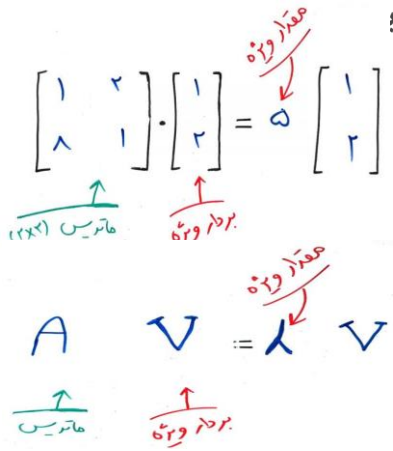
فاصله‌ی میکوفسکی در واقع حالت عمومیِ فاصله‌ی اقلیدسی و فاصله‌ی منهتن است. این فاصله با استفاده از یک پارامتر به عنوان مرتبه **order** می‌تواند فرمول خود را تغییر دهد. فرمول کلی فاصله‌ی مینکوفسکی به صورت زیر است:

در فرمول زیر اگر **p** برابر ۱ باشد، فرمول شبیه به فاصله‌ی منهتن شده و اگر **p** برابر ۲ باشد، فرمول شبیه به فاصله‌ی اقلیدسی می‌شود. در کاربردهای مختلف از فاصله‌ی مینکوفسکی معمولاً با پارامترِ **p** بین ۱ تا ۴ استفاده می‌شود.

$$Distance(p,\ q) = \sqrt[p]{\sum_{i=1}^{n} |q_i - p_i|^p}$$

بردار ویژه ومقدار ویژه برای یک ماتریس

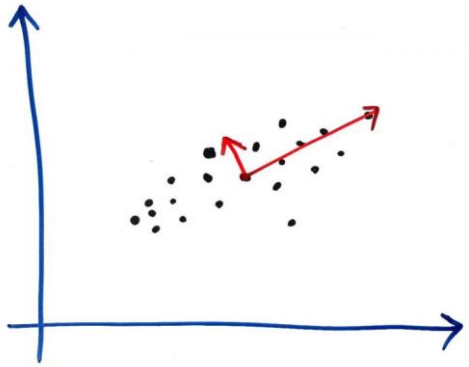
بردار ویژه یا همان **EigenVector**و مقدار ویژه یا همان **EigenValue** دو عنصری هستند که در بخش‌های مختلف یادگیری ماشین و داده‌کاوی کاربرد فراوانی دارند. به ویژه در کاهش ابعاد و مهندسی ویژگی، این دو مقدار می‌توانند کاربردی باشند.مثلا یک ماتریس**A** داریم.بردارِ ویژه برداری است که وقتی در ماتریس ضرب می‌شود، حاصل آن برابر است با ضربِ همان بردارِ ویژه در عددی مانند عدد ۵ که در تصویر مقابل مشخص شده است. تصویر زیر که فرموله شده‌ی همان مثال را نشان می دهد .

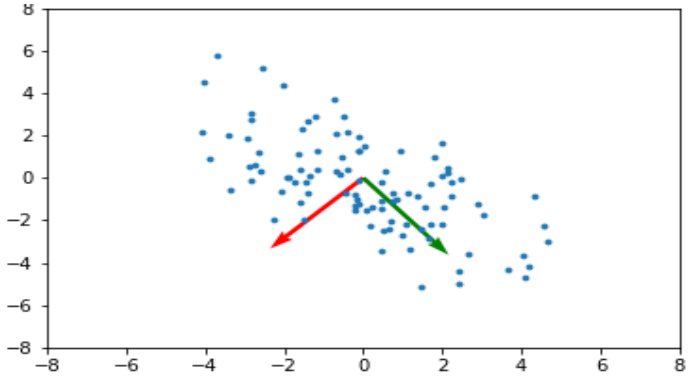


در واقع بردارهای ویژه می‌توانند، جهت توزیع‌شدگیِ داده‌ها را به نوعی مشخص کنند.

نقاطِ سیاه داده‌هایی هستند که بر روی دو بُعد (آبی رنگ) قرار دارند ،دو بردارِ قرمز رنگ **بردارهای ویژه** برای این داده‌ها هستند که به نوعی **جهت توزیع** شدن داده‌ها را مشخص می‌کنند. توجه داشته باشید که به ازای هر بُعد (یعنی هر ستون) یک بردار ویژه‌ی جدید خواهیم داشت. مثلاً اگر مسئله‌ی ما ۳۰ بُعد داشته باشد، تعداد بردارهای ویژه هم برابر با ۳۰ است. انیمیشن زیر می‌تواند در درک بهتر بردار ویژه کمک کند:

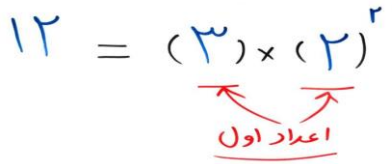
در شکل زیر، چون نمونه‌ها بر روی **دو بُعد** گسترده شده‌اند، دو بردار قرمز و سبز رنگ (که همان بردارهای ویژه هستند) داریم.بردارِ ویژه کاربردِ مختلفی در علوم مهندسی دارد که یکی از کاربردهای آن در الگوریتم یادگیری‌ماشین جهت انجام عملیاتِ کاهشِ ویژگی یا کاهشِ ابعاد است .



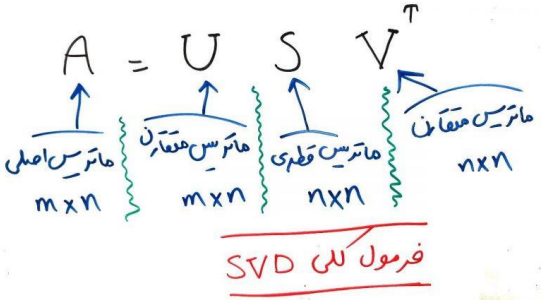


SVDیا همان Singular Value Decomposition در ماتریس چیست؟

با تجزیه‌ی یک عدد به اعداد اول که آشنا هستید؟ مثلا می‌گوییم عدد ۱۲ را می‌خواهیم به اعداد اول تجزیه کنیم. شکل مقابل را مشاهده کنید:

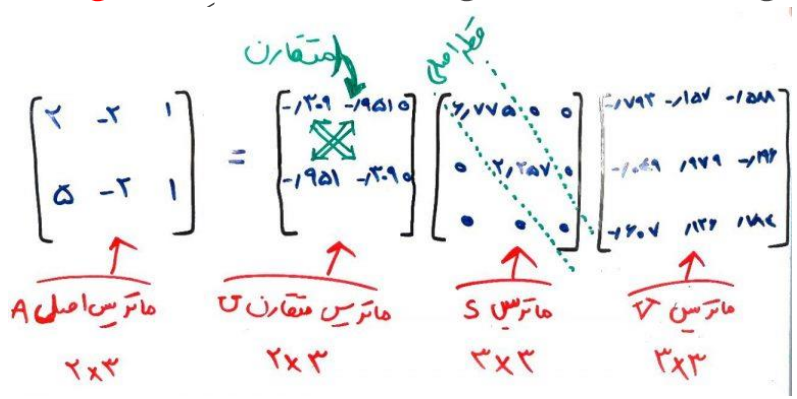


همین‌کار را هم می‌توانیم برای ماتریس‌ها انجام دهیم. یعنی یک ماتریس را به عواملِ سازنده‌ی آن تجزیه کنیم. یکی از روش‌های این تجزیه Singular Value Decompositionیا همان SVDاست ، فرمول زیر، پایه‌ی SVDرا تشکیل می‌دهد:



ماتریس را می‌تواند به سه ماتریس دیگر شکسته شود. ماتریسِ Aیک ماتریسِ mدر nاست. یعنی mسطر دارد و nستون. ماتریس Uیک ماتریسِ mدر nو متفارن است. ماتریسِ Sیک ماتریسِ قُطری diagonalبه صورت nدر nاست (یعنی فقط قطرِ اصلیِ آن عدد دارد و بقیه صفر است). ماتریسِ Vهم یک ماتریس متفارنِ nدر nاست. در واقع ماتریسِ اصلی ماتریس Aتشکیل شده از ضربِ این سه ماتریس Uو Sو Vاست. به این کار در اصلاح تقسیم‌بندی Factorizationمی‌گویند و به این معنی است که ماتریسِ Aمی‌تواند از سه ماتریسِ Uو Sو Vتشکیل شود.

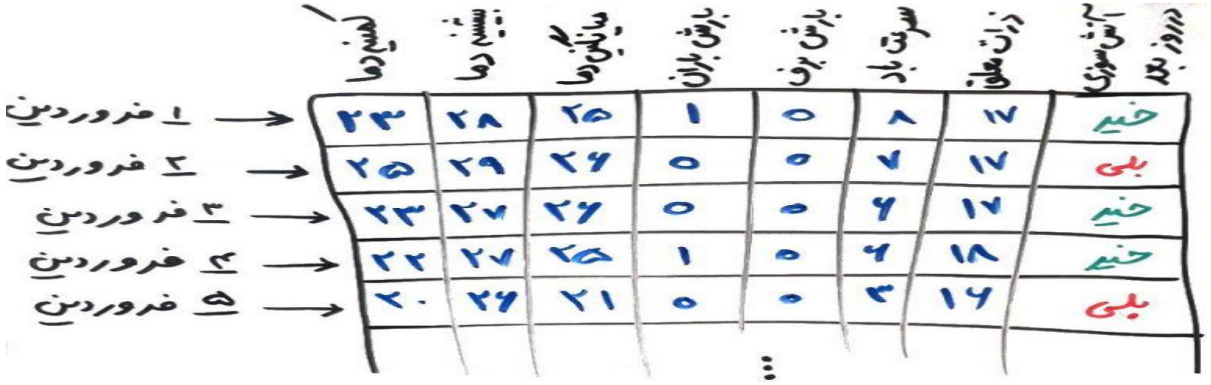
در شکل مقابل نمونه‌ای از تقسیم‌بندی Factorizationماتریسِ Aبه سه ماتریس را نشان داده‌شده است:



SVDیا همان Singular Value Decomposition در ماتریس چیست؟

مقادیری که در S وجود دارد در واقع همان مقادیرِ منحصر به فردی Singularاست که می‌تواند در عملیاتِ کاهشِ ویژگی به ما کمک کند. برای مثال با استفاده از تجزیه‌ی ماتریس به کمک SVDمی‌توانیم متوجه شویم که کدام یک از ستون‌های یک ماتریس، دارای اطلاعات بیشتری هستند.

ما برای ذخیره‌سازی داده‌ها در داده‌کاوی از ماتریس استفاده می‌کردیم. مثلا شکل زیر را از بخش قبل «کاربرد ماتریس‌ها در داده‌کاوی» به یاد بیاورید.همان‌طور که می‌بینید، هر سطر یک روز را نشان می‌دهد و هر ستون یک ویژگی از آن روز. مثلاً در روز اول فروردین، میانگین دمای هوا ۲۵ درجه بوده است. ما می‌خواهیم از روی این ویژگی‌ها متوجه شویم که آیا فردای یک روز، جنگل آتش خواهد گرفت یا خیر (که این کار وظیفه‌ی الگوریتم یادگیری ماشین است). حتماً می‌دانید که هر کدام از این ستون‌ها، برای پیش‌بینیِ اینکه فردا آتش‌سوزی خواهیم داشت یا خیر موثر هستند ولی برخی کمتر و برخی بیشتر تاثیر دارند. یکی از روش‌هایی که می‌توانیم بفهمیم که کدام ستون (ویژگی) تاثیر بیشتری دارد (یعنی اطلاعات بیشتری در خود قرار داده است)، همین روش SVDاست. در واقع SVDمی‌تواند برای مثال این ۷ ویژگی را به صورت فشرده شده به ۳ ویژگی تبدیل کند و به این صورت، حجم ذخیره‌سازی و عملیاتی که الگوریتم باید بر روی آن انجام دهد، به مراتب کمتر می‌شود..البته که محاسبه‌ی SVDبرای یک ماتریسْ دارای پیچیدگیِ زمانی خوبی نیست ولی به هر حال این روش یکی از روش‌های شناخته شده در کاهشِ ویژگی‌ها در داده‌کاوی می‌باشد و کاربردهای دیگرِ آن نیز در قسمتهای بعد آشنا می‌شویم. در زبان‌های برنامه‌نویسی مانند Pythonو Rمی‌توانید کتابخانه‌های مختلفی را پیدا کنید که این عملیات را به سادگی برای شما انجام می‌دهند.



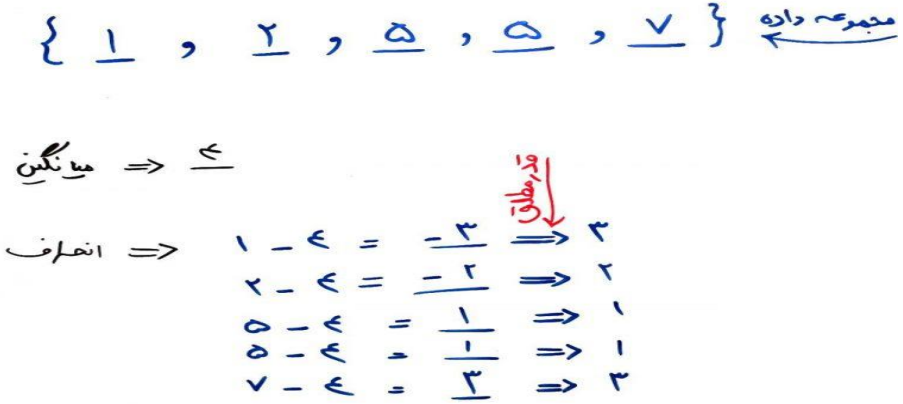
سید ایمان غفوریان - عضو هیات علمی دانشگاه آزاد مشهد-زمستان ۱۴۰۲

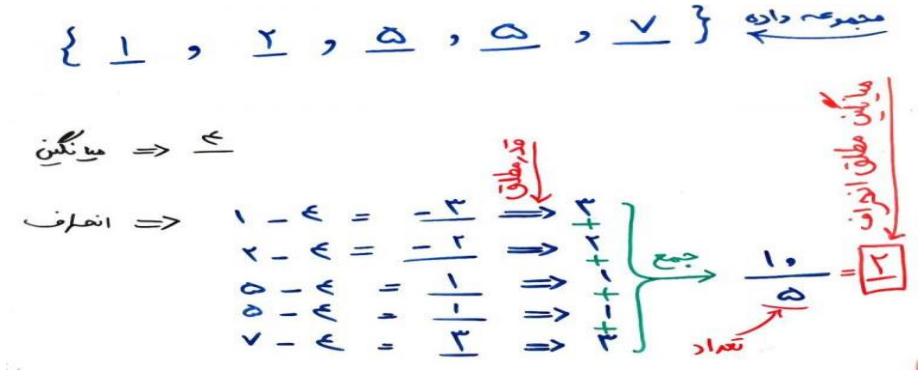
تخمین تنوع و پراکندگی Estimation Of Variabilityو انواع مختلف آن

برای تعریف پراکندگی داده‌ها (یا مجموعه‌ای از یک سری داده‌ها) تعاریف و معیارهای مختلفی موجود است که هر کدام کاربرد خاص خود را دارند. احتمالا اگر به صورت کاربردی یا حتی تئوری علوم داده را خوانده باشید در چند جا به انواع معیارهای پراکندگی برخورده‌اید. چند مورد از مهم‌ترین معیارهای پراکندگی را با هم مرور می کنیم.

داده‌های زیر را فرض کنید: ۱، ۲، ۵، ۵، ۷ میانگین این داده‌ها برابر ۴ است. یعنی فرض کنید تخمین مکان را برای این داده‌ها معیار میانگین در نظر گرفته‌ایم. حال اگر بخواهیم انحراف یا همان Deviationرا تعریف کنیم مانند شکل زیر است:

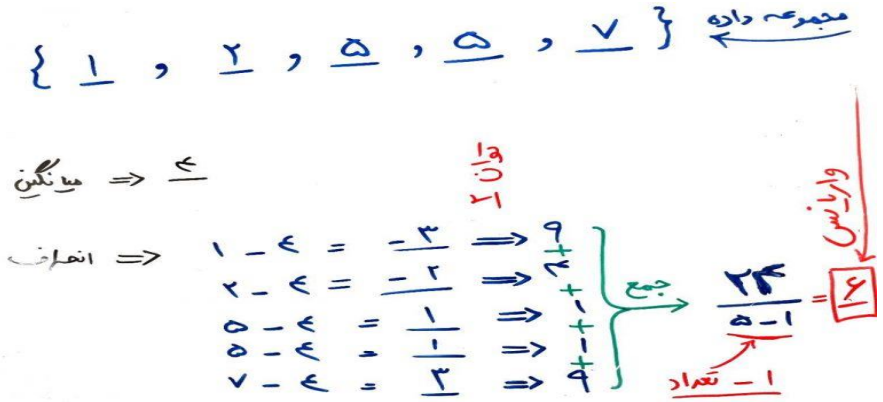
در واقع اختلاف هر کدام از اعداد نسبت به میانگین را اینجا انحراف یا همان Deviationدر نظر گرفته‌ایم. البته برای انحراف می‌توانیم مقدار قدر مطلق یا همان Absolute را در نظر گرفته‌ایم. یعنی مقدار منفی‌ها را قدر مطلق گرفته و مثبت کرده‌ایم. حال اگر میانگین این انحرافات را به دست بیاوریم یک معیار استاندارد را حساب کرده‌ایم. این معیار همان میانگین قدر مطلق انحراف یا انحراف مطلق میانگین یا Mean Absolute Deviationاست. شکل زیر را نگاه کنید:





تخمین تنوع و پراکندگی Estimation Of Variabilityو انواع مختلف آن

البته که Mean Absolute Deviationتنها معیار برای تخمین پراکندگی نیست. معروف‌ترین معیار در این حوزه شاید همان واریانس Varianceباشد. شکل زیر واریانس را برای اعداد بالا محاسبه کرده‌است:

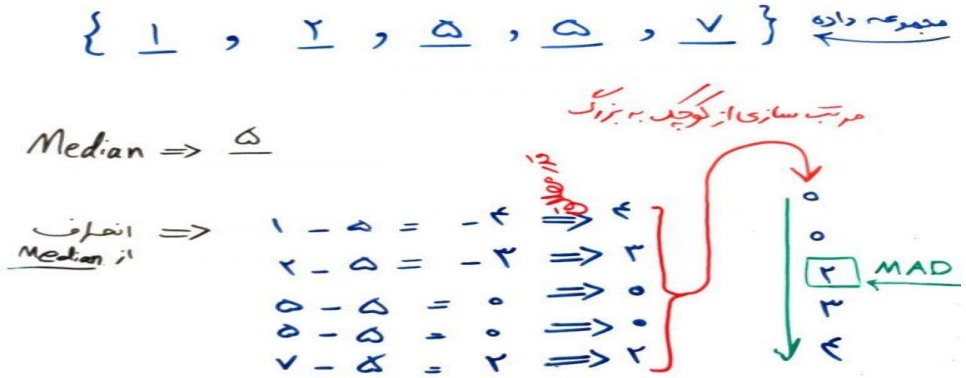


اگر از واریانس رادیکال بگیرید، معیار دیگری به اسم انحراف استاندارد یا همان Standard Deviationبه دست می‌آید که عددی که از انحراف استاندارد به دست می‌آید نسبت به واریانس گویاتر است. زیرا واریانس هر کدام از اختلافات را به توان ۲ رسانده است. **در واقع مثلا وقتی می‌گوییم انحراف استاندارد برابر ۲.۴ است یعنی هر کدام از اعداد تقریبا به اندازه ۲.۴واحد (به طور میانگین) از معیار میانگین مجموعه اعداد فاصله دارند. ولی وقتی بگوییم واریانس برابر ۶است، چیزی دستگیرمان نمی‌شود. (در واقع معیار انحراف استاندارد قابل تفسیرتر است) فقط می‌توانیم واریانس دو مجموعه‌ی مختلف را با هم مقایسه کنیم. یعنی اگر یک مجموعه‌ای واریانس بیشتری نسبت به یک مجموعه‌ی دیگر داشت، می‌توانیم بفهمیم که این مجموعه پراکندگی بیشتری دارد.**

دو معیاری واریانس و انحراف استاندارد نسبت به داده‌های پرت حساس هستند به این معنی که داده‌های پرت می‌توانند بر روی آن‌ها تاثیر زیادی بگذارند.

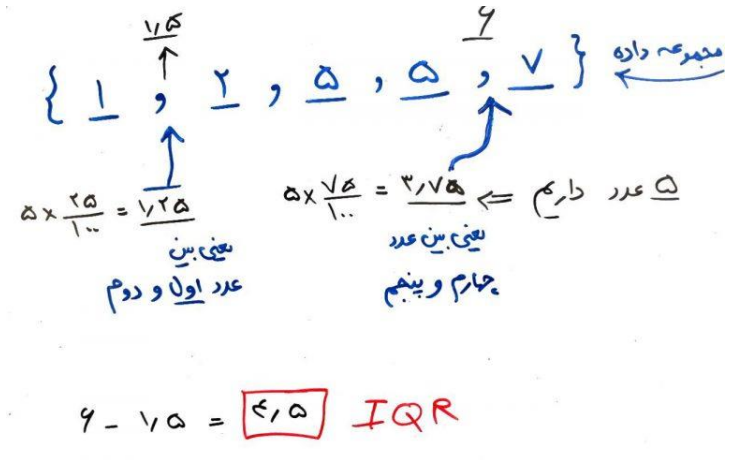
تخمین تنوع و پراکندگی Estimation Of Variabilityو انواع مختلف آن

Medianیکی از معیارهای تخمین مکان بود که نسبت به داده‌های پرت قدرتمند Robustبود. یعنی داده‌های پرت نمی‌توانستند تاثیر زیادی بر روی محاسبه Medianبگذارند. حال از روی همین Medianبه معیاری می‌رسیم که به آن Median Absolute Deviationمی‌گویند. اگر Medianبرای داده‌های بالا ۵ باشد، این معیار مانند شکل زیر محاسبه می‌شود:



تا اینجا معیارهای معروفی را گفتیم، اما معیارهای دیگری در میان تخمین‌های پراکندگی وجود دارد. یک معیار دیگر، معیار بازه یا Rangeاست. معیار ساده‌ای که در آن کافیست اختلاف کمترین و بیشترین مقدار را در یک مجموعه داده محاسبه کنید. برای مثال در داده‌های بالا اختلاف کمترین و بیشترر عدد (۱ – ۷) برابر است با ۶. یعنی معیار Rangeبرای این داده‌ها برابر ۶ است. اما معیار Rangeبسیار حساس به داده‌های پرت است. برای همین یک معیار بهتر به نام IQRکه مخفف InterQuartile Rangeاست معرفی شده است. در این معیار، بازه‌ای بین ۲۵درصد تا ۷۵درصد داده‌ها را محاسبه می‌کنیم. سپس اختلاف بیشتر مقدار باقی‌مانده و کمترین مقدار باقی‌مانده را گرفته تا IQRبرای داده‌ها محاسبه شود. این معیار نسبت به داده‌های پرت مقاومت بیشتری دارد. شکل زیر IQRرا برای داده‌های بالا محاسبه کرده است .

تخمین تنوع و پراکندگی Estimation Of Variability و انواع مختلف آن



طبیعتا معیارها و سنجه‌های مختلف دیگری را هم می‌توان برای پراکندگی داده‌ها محاسبه کرد که با کمی فکر کردن می‌توانید معیاری که برای داده‌ها و نوع داده‌های شما مناسب باشد را پیدا کنید.

ماتریس کواریانس چیست؟(Covariance)و ماتریس همبستگی (Correlation)

ماتریسِ همبستگی یا همان Correlation Matrixرا نمایش دهیم. در این ماتریس متغیرهای همان ویژگی‌های مجموعه‌ی داده هستند. برای مثال یک سری پستاندار را می‌خواهیم مورد بررسی قرار دهیم. در این بررسی برای هر پستاندار ۳ویژگی در نظر می‌گیریم. وزن، ساعت خواب و طول عمر حالا شکل زیر را ببینید ، این یک ماتریس است که ۳سطر و ۳ستون دارد و متقارن است. توجه کنید که تعدادِ سطر و ستون‌ها برابرِ تعداد ویژگی‌های مجموعه‌ی داده (در این‌جا ۳) است. سطرها و ستون‌های این ماتریس برابرند. هر کدام از خانه‌ها با عددی مشخص شده‌اند که در بازه‌ی منفی ۱ تا مثبت ۱ قرار دارند. هر چه این عدد کمتر باشد به این معنی است که دو ویژگی (در محل تقاطع آن عدد) به همدیگر ارتباط معکوس دارند و هر چه قدر این عدد بزرگتر باشد یعنی دو ویژگی به همدیگر وابستگیِ مثبت دارند. برای درک بهتر، عددی که در شکلِ بالا سبز رنگ کردیم را مشاهده کنید. عدد منفی ۰/۳۰۷ به این معنی است که در بین این گونه پستانداران با زیاد شدن وزن آن‌ها، ساعات خوابشان کمتر می‌شود. یعنی دو ویژگیِ وزن و ساعت خواب به همدیگر به اندازه ۰/۳۰۷ وابستگی منفی دارند. حالا عددی که با رنگ قرمز مشخص شده را مشاهده کنید. همان‌طور که می‌بینید، این عدد در نقطه‌ی تقاطعِ دو ویژگیِ طول عمر و وزن قرار دارد و به خاطر مثبت بودنْ نشان می‌دهد که هر چه وزنِ یک پستاندار بیشتر باشد، طولِ عمر او نیز بیشتر می‌شود. در واقع این دو متغیر به اندازه‌ی ۰/۳۰۲ به همدیگر وابستگیِ مثبت دارند. قطعاً توجه دارید که قُطرِ اصلیِ این ماتریس برابرِ ۱ هست زیرا هر ویژگی با خودش طبیعتاً همبستگی حداکثری دارد.

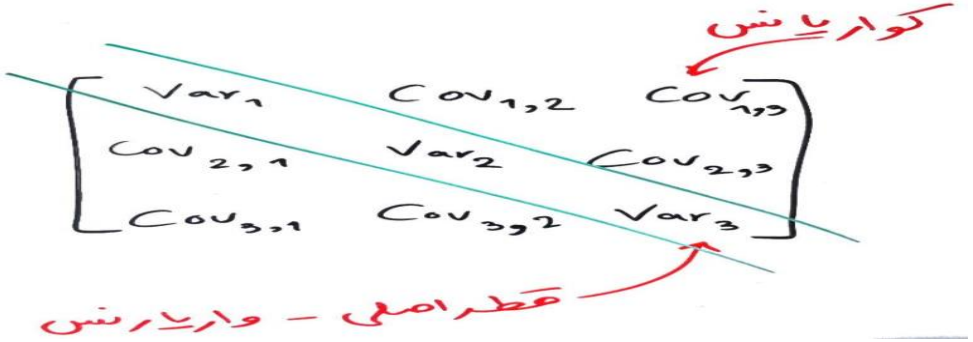
	وزن	ساعت خواب	طول عمر
وزن	۱	-۰/۳۰۷	۰/۳۰۲
ساعت خواب	-۰/۳۰۷	۱	-۰/۴۱۰
طول عمر	۰/۳۰۲	-۰/۴۱۰	۱

ماتریس کواریانس چیست؟(Covariance)و ماتریس همبستگی (Correlation)

کواریانس :این اعداد و ویژگی‌هایی که در موردِ آن‌ها بحث گردید ، مقدارِ همبستگیِ دو ویژگی (دو متغیر) را نشان می‌دهد.در بعضی از مراجع از کواریانس Covariance نیز برای این رابطه نام برده می‌شود. کواریانس در واقع یک حالتِ غیرنرمال (غیر استاندارد) از همبستگی Correlationاست. زیرا برای محاسبه‌ی همبستگی باید مقدارِ کواریانسِ بین دو ویژگی را تقسیم بر انحرافِ استانداردِ (انحراف معیار) دو متغیر کرد ، این کار (تقسیم بر انحراف استاندارد) باعث می‌شود مقدارِ اعداد در بازه‌ی منفیِ ۱ تا مثبتِ ۱ قرار بگیرند و بتوان آن‌ها را با هم مقایسه کرد.

زیرا مقدار کواریانس در بازه‌ی منفی ۱ و مثبت ۱ نیست و باتوجه به دامنه تغییرات یک ویژگی می‌تواند خیلی زیاد یا خیلی کم شود. در مثالِ بالا، مقدارِ سن ممکن است بین ۵ تا ۱۰۰متغیر باشد ولی مقدار وزن می‌تواند بین ۰/۵کیلوگرم تا ۵۰۰کیلوگرم در بین پستانداران باشد و این دامنه‌ی تغییرات بر روی مقادیرِ کواریانس اثر می‌گذارد و مانع از مقایسه درست اعداد داخل ماتریس نسبت به هم می‌شود. حال برای فهم ماتریس کواریانس به ماتریس زیر نگاه کنید:

واریانس دامنه‌ی تغییرات یک متغیر نسبت به خودش است. در حالی که کواریانس دامنه‌ی تغییرات دو متغیر نسبت به همدیگر است. یعنی به نوعی، پاسخ به این سوال است که مثلا با کم شدن مقدار یک ویژگی (مانند سن پستانداران)، ویژگی دیگر (مانند وزن پستاندار) چه تغییری پیدا می‌کند؟ و کواریانس هر ویژگی با خودش همان واریانس Varianceآن ویژگی است.



آنالیز مولفه اصلی PCA– Principal Component Analysis یا همان چیست؟

در ادامه به یکی از مباحثِ اصلی و پیشرفته‌تر در جبر خطی که به آن آنالیز مولفه اصلی Principal Component Analysisیا به اختصار PCAمی‌گویند و همچنین کاربرد آن را در مسائلِ حوزه‌ی علوم‌داده Data Scienceمی پردازیم.

یکی از کاربردهای اصلیِ PCAدر عملیاتِ کاهشِ ویژگی imensionality Reductionاست. PCAهمان‌طور که از نامش پیداست می‌تواند مولفه‌های اصلی را شناسایی کند و به ما کمک می‌کند تا به جای اینکه تمامیِ ویژگی‌ها را مورد بررسی قرار دهیم، یک سری ویژگی‌هایی را ارزشِ بیشتری دارند، تحلیل کنیم. در واقع PCAآن ویژگی‌هایی را که ارزش بیشتری فراهم می‌کنند برای ما استخراج می‌کند .

فرض کنید یک فروشگاه می‌خواهد ببیند که رفتار مشتریانش در خریدِ یک محصول خاص (مثلا یک کفش خاص) چطور بوده است. این فروشگاه، اطلاعات زیادی از هر فرد دارد (همان ویژگی‌های آن فرد). برای مثال این فروشگاه، از هر مشتری ویژگی‌های زیر را جمع‌آوری کرده است:

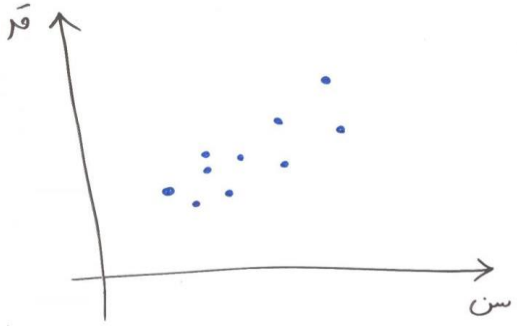
سن، قد، جنسیت، محل تولد شخص (غرب ایران، شمال ایران، شرق ایران یا جنوب ایران)، میانگین تعداد افراد خانواده، میانگین درآمد، اتومبیل شخصی دارد یا خیر و در نهایت اینکه این شخص بعد از بازدید کفش خریده است یا خیر. ۷ویژگیِ اول ابعاد مسئله ما را می‌ساختند و ویژگیِ آخر هدف Target می‌باشد . PCAمی‌تواند آن مولفه‌هایی را انتخاب کند که نقش مهمتری در خرید دارند. برای مثال، مدیرِ فروش به ما گفته است که به جای اینکه هر ۷بُعد را در تصمیم‌گیری دخالت دهیم، نیاز به ۳بُعد (۳ویژگی) داریم تا بتوانیم آن‌ها را بر روی یک رابطِ گرافیکی ۳بُعدی به نمایش دربیاوریم. پس در واقع نیاز داریم ۷بُعد را به ۳بُعد کاهش دهیم. به این کار در اصطلاح کاهش ابعاد یا Dimensionality Reductionمی‌گویند. PCAمی‌تواند این کار را برای ما انجام دهد.

آنالیز مولفه اصلی PCA– Principal Component Analysis یا همان چیست؟

PCAمی‌تواند این کار را برای ما انجام دهد. PCAبا توجه داده‌ها و دامنه‌ی تغییراتِ هر کدام از آن‌ها، می‌تواند ویژگی‌هایی را انتخاب کند که تاثیر حداکثری در نتیجه نهایی داشته باشند. در مثال بالا (فروشگاه)، فرض کنید ویژگی قد، تاثیر زیادی در اینکه یک فرد از فروشگاه خرید کند نداشته باشد. PCAاین قضیه را متوجه می‌شود و در الگوریتمِ خود ویژگیِ قد را تا جای ممکن حذف می‌کند. در واقع در فرآیند تبدیلِ ۷ویژگی به ۳ویژگیِ نهایی (که با توجه به درخواست مدیرِ فروش به دنبال آن هستیم) PCA ویژگیِ قد را کمتر دخالت می‌دهد. این‌گونه است که ویژگی‌های مهمتر از نظر PCAوزن بیشتری در تولیدِ ویژگی‌های کاهش یافته پیدا می‌کنند.

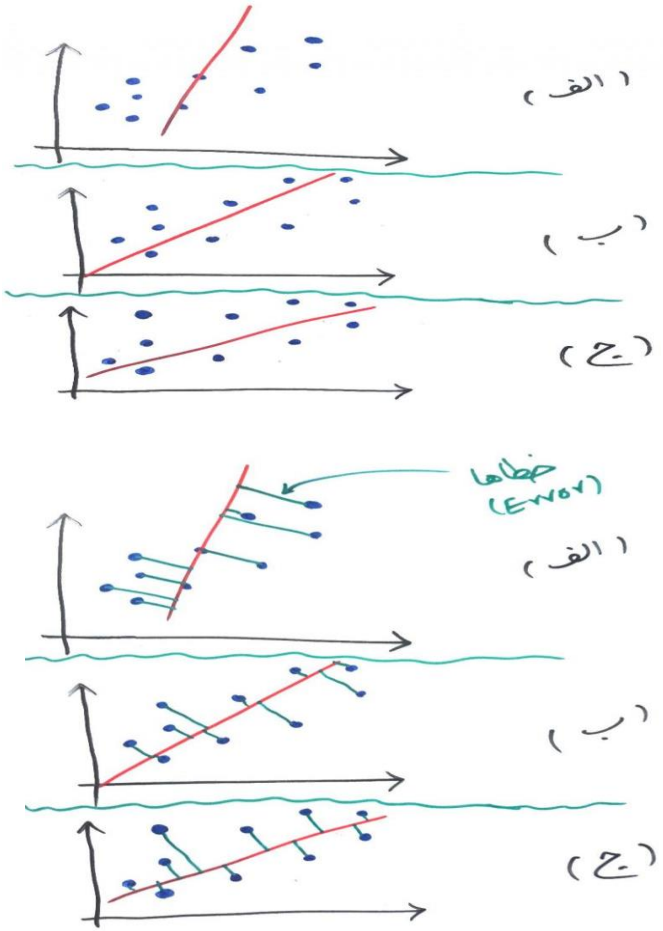
البته این بدان معنا نیست که در فرآیند کاهش ابعاد، PCAدقیقا همان ویژگی را حذف می‌کند. بلکه PCAتوان این را دارد که به یک سری ویژگی جدید برسد. مثلا این الگوریتم ممکن است به این نتیجه برسد که افرادی که در شمال و غرب ایران زندگی می‌کنند و سنِ آن‌ها بالای ۴۰سال است، احتمال خرید بالایی دارند در حالی که برعکس این قضیه احتمال خرید را بسیار کمتر می‌کند. در واقع این‌جا PCAبه یک ویژگی ترکیبی از محل تولد شخص و سن رسیده است. این دقیقا یکی از قدرت‌های الگوریتم PCAدر کار بر روی داده‌ها است.

برای سادگی فرض کنید داده‌های مشتریان ما کلا ۲بُعد دارند. سن و قد. حال آن‌ها را بر روی محور مختصات نمایش می‌دهیم. فرض کنید شکلی مانند شکل زیر تشکیل می‌شود:



هر کدام از نقاط آبی رنگ ، در مثال ما یک مشتری است که با توجه به ویژگی سن (محور افقی) و ویژگی قد (محور عمودی) در صفحه مختصات رسم شده است. همان‌طور که از درس بردار ویژه Eigen Vector به یاد دارید، بردار ویژه می‌تواند کمک کند تا در میان داده‌های ما خطی کشیده شود که بیشترین دامنه تغییرات در امتداد آن خط رخ داده باشد. حال به تصویر زیر نگاه کنید:

آنالیز مولفه اصلی PCA– Principal Component Analysis یا همان چیست؟

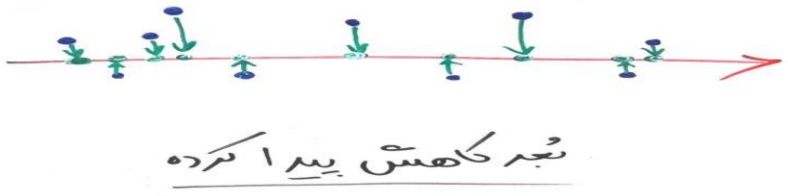


همان‌طور که بیان شد هر کدام از نقاط آبی رنگ، در مثالِ ما یک مشتری است که با توجه به ویژگی سن (محور افقی) و ویژگی قد (محور عمودی) در صفحه مختصات رسم شده است. فاصله هر نقطه آبی تا خط قرمز را می‌توان به عنوان یک خطا Errorدر نظر گرفت و به تبعِ آن مجموعِ خطا برابر است با جمع فاصله تک تکِ نقاط آبی تا خط قرمز. به شکل زیر نگاه کنید، کدام تصویر (الف، ب یا ج) مجموع خطاهای کمتری دارند؟

فاصله نقاط نسبت به خط قرمز با خط با رنگ سبز مشخص شده‌اند. اگر جمع این فاصله خطا – Error را برای هر نمودار برابر خطای کلی داده‌ها نسبت به خط قرمز در نظر بگیریم، تصویرِ الف بیشترین میزان خطا را دارد و بعد از آن تصویر ب و در نهایت تصویر ج کمترین خطا را دارد. PCA_به دنبال ساختنِ خطی مانند خط ج است (که در واقع همان بردار ویژه ماست) که کمترین خطا Least Errorرا داشته باشد. با این‌کار هر کدام از نقاط بر روی خط قرمز نگاشت می‌شوند و در تصویرِ مقابل که ۲بُعدی است می‌توان این ۲بُعد را به ۱بُعد (که همان خط قرمز رنگ است) نگاشت کرد.

آنالیز مولفه اصلی (PCA–Principal Component Analysis یا همان چیست؟

در نهایت می‌تواند چیزی مانند شکل مقابل رخ دهد:



در این مثال آخر ما ۲بُعد را به ۱بُعد کاهش دادیم. البته در مثال‌های واقعی ممکن است ۱۰۰۰بُعد را به ۲بُعد کاهش دهند تا بتوان آن را بر روی یک نمودار به نمایش درآورد و این کار با با PCAانجام دهند که هم از سرعتِ معقولی برخوردار است و هم کیفیت قابل قبولی دارد.