Applications of SVD in Data Science

Parisa Ghanad Torshizi

یکی از انواع تجزیه ی ماتریس، تجزیه ی SVD نام دارد که در آن هر ماتریس A به صورت حاصلضرب سه ماتریس دیگر نوشته میشود. اگر فرض کنیم ماتریس A از رتبه ی r باشد آنگاه می توانیم آن را به صورت مجموع r تا ماتریس به شکل $\sigma_i u_i v_i$ بنویسیم به طوری هرکدام از این ماتریس ها از مرتبه ی ۱ می باشند.

$$A = U\Sigma V^T = \sigma_1 u_1 v_1 + \sigma_2 u_2 v_2 + \sigma_3 u_3 v_3 + \cdots + \sigma_r u_r v_r$$

هر کدام از این ماتریس ها حاوی اطلاعاتی از ماتریس اصلی می باشند .از آنجایی که در تجزیه ی SVD، ترتیب مقادیر منفرد به این گونه است: $\sigma_1 > \sigma_2 > \cdots > \sigma_r$ و ماتریس های u وv به علت متعامد بودن اندازه ی یک دارند. پس ماتریس های اولی ،بزرگ ترند و اطلاعات بیشتری از ماتریس اصلی را با خود دارند. بنابرهمین خصوصیت میتوان از ماتریس های کوچک صرف نظر کرد و تنها k تا مقادیر منفرد اولی و یعنی k تا ماتریس اولی را در نظر گرفت. پس خواهیم داشت:

$$A_k = U_k \Sigma_k V_k^T = \sigma_1 u_1 v_1 + \sigma_2 u_2 v_2 + \sigma_3 u_3 v_3 + \cdots + \sigma_k u_k v_k$$

ما به دنبال k بهینه هستیم به طوریکه A_k تقریب خوبی از A باشد.

$$\min_k \|A - A_k\|$$

از این خاصیت می توان در بسیاری از موارد علوم داده بهره برد:

1) Image compression :

از این شیوه برای کاهش ابعاد یک تصویر می توان استفاده کرد. ابتدا تصویر را به صورت ماتریس در می آوریم. سپس با انتخاب مقدار k تصویر را با کمک تجزیه SVDدوباره میسازیم.

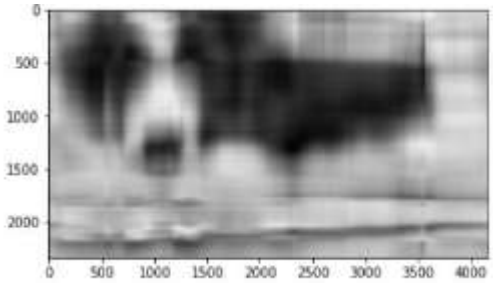
هرچه k انتخاب شده بزرگ تر باشد تصویر ساخته شده به تصویر اصلی نزدیک تر خواهد بود.

تصویر زیر را در نظر بگیرید:

با انتخاب k های مختلف خواهیم داشت:

k=30

K=5

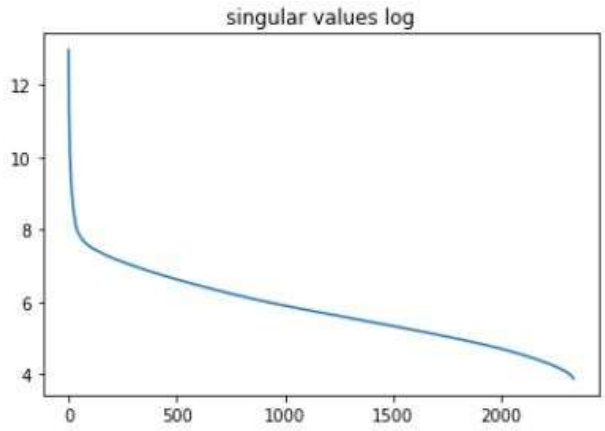


با انتخاب k=5 محتوای کلی شکل حفظ شده ولی جزئیات نه.

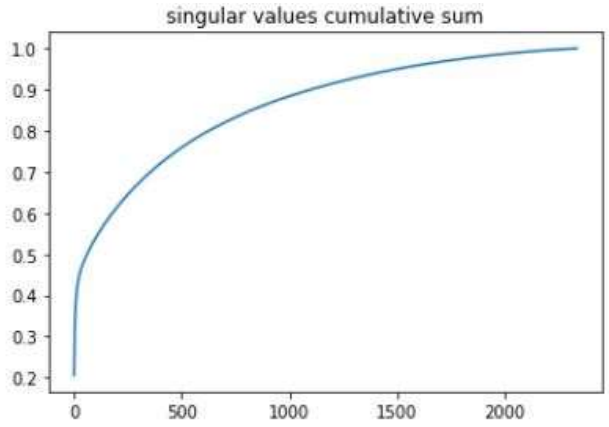
برای ذخیره ی ماتریس اصلی به 2336*4160=9717760 فضا نیاز داریم درحالی که با انتخاب k=30 به فضای (1+2336+4160)*30=194910 نیاز است یعنی تنها ۲ درصد فضای مورد نیاز برای فضای ماتریس اصلی!

از آنجایی که تعداد زیادی از ستون های ماتریس تصویر همبستگی دارند، نیازی به ذخیره ی همه ی آنها نیست در نتیجه با تعداد کمتری از مقادیر ویژه می توان تصویر را با دقت نسبتا خوبی ساخت.

نمودار لگاریتم مقادیر منفرد را رسم میکنیم .همانطور که انتظار می رود به ترتیب مقادیر ویژه کاهش می یابند.



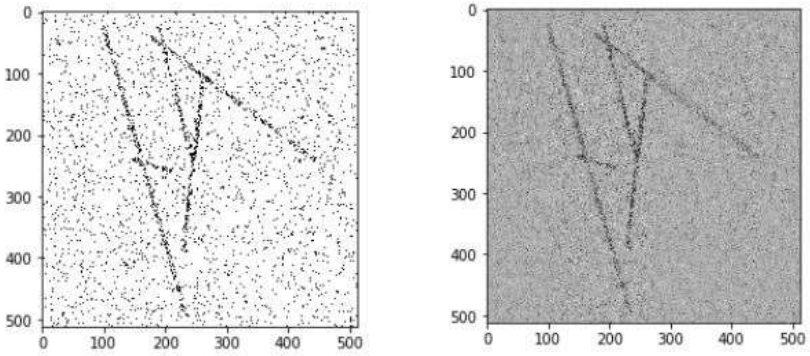
نمودار دوم نسبت مجموع مقادیر منفرد تا مقدارمنفرد kام به مجموع کل مقادیر ویژه را نشان می دهد. این نمودار به ما می گوید که با ذخیره کردن k تا مقدار منفرد چه میزان از انرژی یا اطلاعات ماتریس اصلی حفظ می شود.



همانطور که می بینیم با افزایش مقدار k ، این نسبت افزایش می یابد. به عنوان مثال با k=200، 60 درصد شکل حفظ شده است.

2) noise reduction :

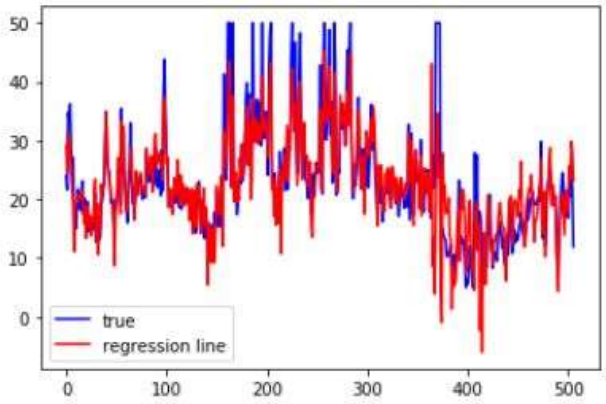
اطلاعات کم اهمیت تر در مقادیر منفرد کوچک تر ذخیره می شوند در نتیجه با نادیده گرفتن آن مقادیر، نویزها از تصویر ساخته شده حذف می شوند.
در شکل زیر تصویر سمت راست بازسازی شده ی تصویر چپ با 70 مقدار منفرد اول آن است. و شکل اصلی در تصویرتا حد زیادی حفظ شده و نویز ها حذف شده اند.



3) Linear regression :

اگر ماتریس خصوصیات را A و بردار هدف را b بنامیم. دنبال ترکیب خطی از A هستیم که تا حد ممکن به بردار هدف نزدیک باشد.این ترکیب خطی توسط بردار x نمایش داده می شود.مساله به حل دستگاه غیرمربعی $AX = b$ تبدیل شد.خواهیم داشت:

$$x = A^{\dagger}b = V\Sigma U^{T}b$$



با انجام مراحل بالا روی ماتریس دادها هی بیماری دیابت، نمودار قرمز به دست می آید که تقریب خوبی از داده های اصلی (نمودار آبی) می باشد.

4) Pca :

از این روش برای کاهش ابعاد دیتا و تصویرکردن دیتا بر روی ابعاد کمتر استفاده میشود .در pca به دنبال این هستیم که بیشترین واریانس دیتا را حفظ کنیم.

موارد استفاده :

- Feature extraction : استخراج اطلاعات و الگوهای مهم از دیتا
- Visualization : توانایی رسم دیتا (رسم دیتای بیشتر از سه بعد عملا غیرممکن است ولی با روش کاهش بعد می توان دیتا را در کمتر از سه بعد رسم کرد.)
- Compression

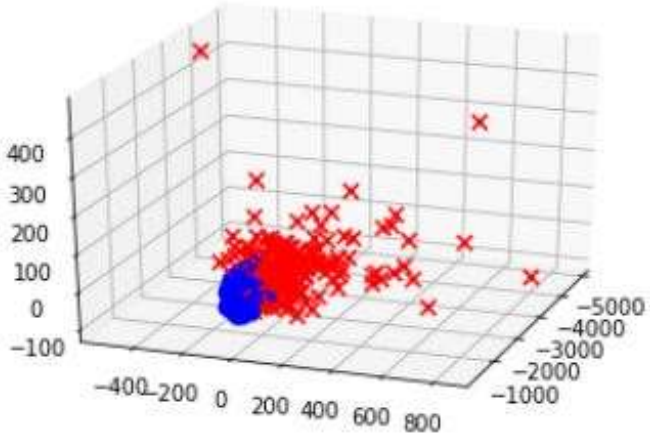
مراحل استفاده ازSVD درPCA :

- محاسبه ی ماتریس میانگین از روی ماتریس اصلی. $\bar{A}$
- $B = A - \bar{A}$
- اعمال SVD بر B و محاسبه ی V
- $T = BV = U\Sigma$

در آزمایش اول میبینیم تجزیه ی SVD شامل تغییر جهت (توسط U) و تغییر اندازه (Σ) می باشد. ابتدا دیتایی شامل 1000 نقطه که توزیع گوسی دارند را درنظر می گیریم و با مقادیر دلخواه مولفه های اصلی نقاط را چرخانده و اندازه آنها را تغییر میدهیم.حال اگر نقاط جدید به دست آمده را تجزیه کنیم می بینیم مقادیر به دست آمده از Σ. U به ترتیب با میزان چرخش و تغییر برابرند.

در آزمایش دوم دیتاست مربوط به سرطان را داریم .که شامل دو دسته ی خوش خیم و بدخیم است.توسط SVD آن را تجزیه میکنیم و براساس فرمول بالا سه مولفه ی اصلی استخراج میکنیم و دیتا بر اساس ان رسم میکنیم .می بینیم دیتا به خوبی دسته بندی شده است.

سه مولفه در راستای محور x و y وz هستند.



دقت داریم که دیتای اولیه همبستگی زیادی دارد .به همین جهت ما مولفه های اصلی را طوری میسازیم که باهم کمترین همبستگی را داشته باشند تا بتوانند واریانس زیادی را پوشش دهند.

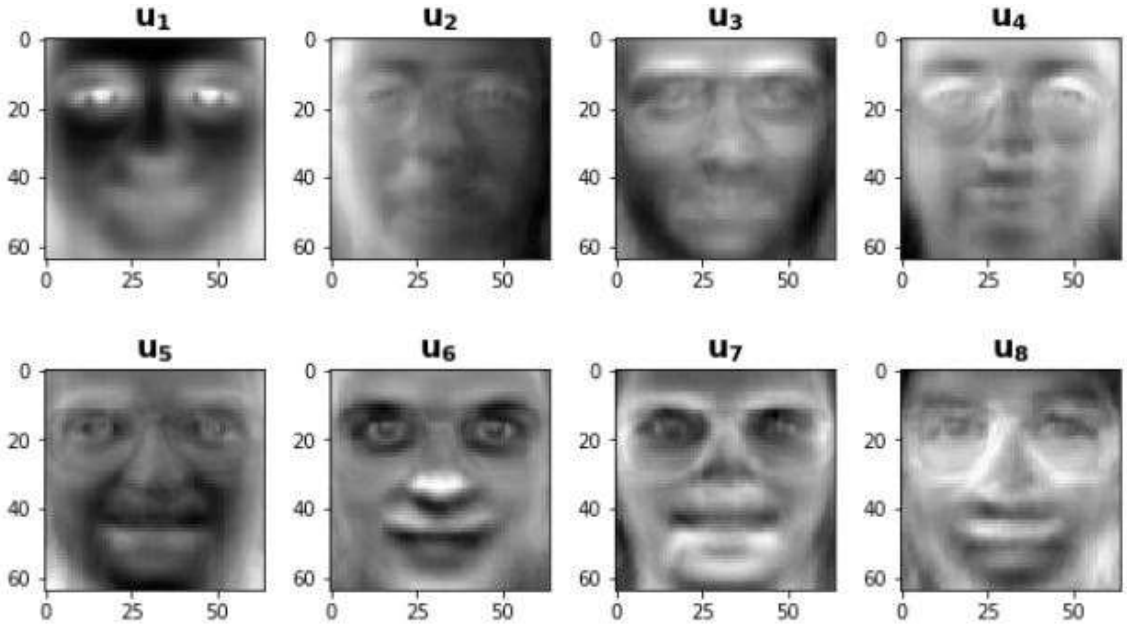
5) Eignfaces :

در آزمایشی ۴۰ نفر انتخاب شده و ازصورت آنها در شرایط مختلف عکس گرفته ایم. ماتریس مربوطه به هر تصویر را تبدیل به بردار کرده و آنها را کنار هم می گذاریم تا ماتریس M که همه ی تصاویر را شامل می شود به دست آوریم.

بردار x را طوری می سازیم به طوری که همه ی درایه های آن به جز درایه ی iام صفر و درایه ی iام ۱ باشند.

M*x به ما تصویر مسطح شده ی iام را می دهد.

در قسمت های قبل دیدیم که ماتریس u به دست آمده ازتجزیه ، اطلاعات مهم ماتریس اصلی را نگهداری میکند. در این جا هم هر ستون u یک تصویر به ما می دهد که به آن eignface می گوییم.هرکدام ازین eignface ها یک مولفه ی تصویر صورت انسان را به ما می دهد.همانطور که می دانیم با توجه به ترتیب مقادیر منفرد، ستون های u هم از اول به اخر اهمیتشان کاهش می یابد. Eignface های دیتاست را در زیر می بینیم :



چشم ها به عنوان مهمترین عامل شناسایی صورت شناخته می شوند. در این جا هم اولین ستون u به چشم ها اختصاص یافته است.

K تا eignface مهمتر را انتخاب میکنیم و آن را U_k می نامیم.

حال وقتی تصویر صورت یک شخص را به ما بدهند آن تصویر را در U_k^T ضرب می کنیم .بردار به دست آمده که α نام دارد، درحقیقت برای آن شخص مثل اثرانگشت عمل میکند که به ما می گوید تصویر این شخص از چگونگی ترکیب خطی این U_i ها به دست می آید.(یعنی این eignface ها با چه ضریبی با هم جمع شده اند.)

تصویر تقریبی را به ازای مقادیر مختلف k با SVD به دست می آوریم.طبیعتا هرچه مقدار k انتخاب شده بیشتر باشد ،تصویر به صورت اصلی شخص نزدیک تر است زیرا جزئیات بیشتری را همراه خود دارد.

یکی از نکات مهم انتخاب k است. برای این کار می توان ماتریس را به صورت مجموع دو ماتریس نویز و ماتریس اصلی نوشت.

$$A = A_{true} + A_{noise}$$

حال اگر نمودار مقادیر منفرد این دو زیرماتریس را روی یک نمودار رسم و سپس مقایسه کنیم. میبینیم مقادیر منفرد ماتریس اصلی از یک جایی به قبل از مقادیر منفرد ماتریس نویز بزرگ ترند.آن محل همان k بهینه است.