جلسه 12

PCA(principal component analysis) : در این حالت unsupervised هست و نباید به صورت پیش بینی به مساله نگاه کرد. هدف اصلی پیش بینی نیست ولی میتوان از خروجی برای پیش بینی استفاده کرد.(تحلیل مولفه های اصلی)

در unsupervised ها هدف بررسی ساختار دادهاست و نه لزوما پیش بینی .

یه مبحث وجود داره به نام cursive dimensionality که بیان میکنه بعدهای زیاد باعث میشود ما نتونیم prediction خوب داشته باشیم و حتی در unsupervisedها نیز تحلیل خوبی داشته باشیم.در تئوری تعداد زیاد متغییر خطا هاور بالاست اما لزوما همیشه اینگونه نیست.(high dimensionality)

این الگوریتم به این صورت کار میکنه که یه دیتا ست میگیزه به ابعاد n\*pو یه خروجی برمیگردونه به ابعاد n\*q که خروجی این الگوریتم هستش . به feature selection معروف است.

PCA همانند SVD میتونه کمک کنه که noise reduction رو انجام بدیم.

Pcaدر کل یک نوع transformation خطی به حساب میاد.

اگر دیتایی داشته باشیم که correlation نداشته باشند این روش زیاد به درد بخور نیست.در صورتی این متد به درد میخوره که این داده ها همپوشانی داشته باشند.

یه نکته که باید بهش خیلی توجه بشه اینه که اگر دوتا خط بر هم عمود باشن Orthogonal هستن و استقلال دارن به این معنی که درPCA اگر دو محور عمود برهم نتخاب کنیم اگر عمود برهم باشن یه اطلاعاتی که میگیریم خیلی کاملا از هم مستقل هستند.

هر چی واریانس بالاتر باشه اطلاعات بیشتری رو از داده ها کاور میکنه به همین دلیل در PCA تلاش میکنیم که واریانس رو ماکزیمم کنیم.

در فرمول مذکور wها همون راستاهای ما خواهند بود.

تفاوت اسکیل متغییرها در PCA تاثیر گذاراست.

به تعداد متغییرها ما راستا خواهیم داشت.مثلا هشت متغییر داریم که 8 راستا خواهیم داشت.

از PCA برای تفسیر هم استفاده خواهد شد. تفسیرهایی که pca به ما میدهد در خیلی از موارد میتواند خیلی بدیهی باشه و در خیلی موارد میتونه اطلاعات خیلی عالی درمورد روند اطلاعات به ما بدهد.

معمولا در pca ستون اول یک overall mean است و ستون دوم یک contrast هستش.

برای مشخص کردن انتخاب چند عدد pca میتوان از دو راه استفاده کرد 1- فرمول: در این حالت واریانس y برابر lambda. درپایتون میتوانیم از p.explained\_variance() استفاده کرد.اولیش واریانس y1 دومیش واریانس y2 و .... که این مقادیر همون eigenvalue هستند. در ادامه میتوان از p.explained\_variance\_ratio\_.consume() میزان متغییرهای و تغییرپذیریها رو دیدو انتخاب کرد.بهترین حالت انتخاب چند تا ستون اول هستش مانند انتخابK تا سیگنال اول درsvd اینجا هم از چن تا سیگنال اول که اصلی هستنداستفاده خواهیم کرد. 2-visualization در این حالت گفته میشود یک bar chart رسم کن از ... و هر جا یک افت زیاد داشتیم به تعداد همونها pca در نظر بگیر

در pca معمولا گفته میشود که اگر در ستونها تفاوت scale زیاد بود باید scala بشن یعنی استناندارد کردن در این حالت یعنی هر ستون رو منهای میانگین همون ستون بکن بر انحراف معیار تقسیم کن که درپایتون با scala انجام میشود. که در sklearn.preprocessing هست استفاده کرد.