Principal Component Analysis

As the no. of features or dimensions in a data set ↑ses ; the amount of data required to obtain a statistically significant result ↑ses.

This results in issues like overfitting, increased computation time, and reduced accuracy.

As the no. of dimensions ↑ses, the no. of possible combination of features ↑ses exponentially ; Making it computationally difficult to obtain a representative sample of the data and it becomes expensive to perform tasks such as clustering or classification.

In some cases, one need more data to achieve the same level of accuracy a lower-dimensional data

This is known as "Curse of dimensionality".

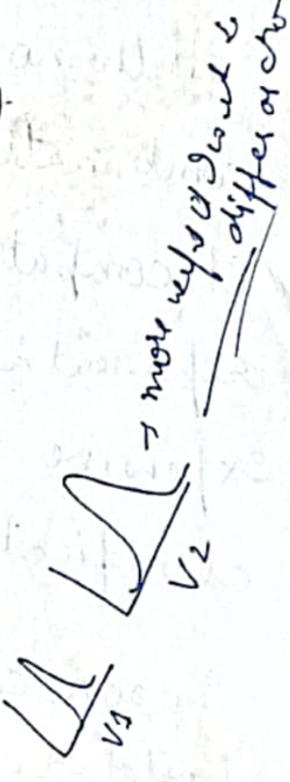
PCA :- Principal Component Analysis is a statistical procedure that uses orthogonal transformation that converts a set of correlated variables to a set of uncorrelated variables : Widely used for dimensionality reduction, lossy data compression, feature extraction + data visualization. It is also known as Karhunen-Loève transformation

PCA is a unsupervised learning algorithm that is used to examine interrelations among a set of variables.

It is also known as a general factor analysis where regression determine a line of best fit

Let say a student ~~has~~ want admission to a grad school. Look at survey

	Academecian	Transport	Safety	Placement
G1	16	12	18	14
G2	12	14	17	20
G3	10	14	11	10
G4	12	14	16	14
G5	15	14	17	19
G6	16	13	17	15
G7	10	14	16	18



$G_1 = \{16, 12, 18, 14\} \in \mathbb{R}^4 \rightarrow$

Wants to reduce to 3 feature y_1, y_2, y_3

$$\mathbb{R}^4 \rightarrow \mathbb{R}^3$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$$

$$y_1 = a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4$$

and so on

,,→ linear combination

we just need to find ~~rh~~ which convert ones

4D into 3D. in such a way that it

preserve the max. information

PCA does this kind of work.

Easy way is that column with less variation is

not that much useful for our decision

Mean:- $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$

Standard deviation

σ :- $= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu)^2}$

Covariance

It is a measure of how two variables change

together

$$\Sigma_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu_x)(y_i-\mu_y)^T$$

Tell us about the relationship b/w multiple variables

in a data set.

...ance mtx

Tell us about which direction it is maximum.

Diagonal element → represents variance of each variable.

Variance tell us how spread out the values of a single variable are from the mean.

off diagonal → covariance b/w pairs of variables.

for a k dimensional dataset $\{X_1, X_2, \ldots, X_k\}$ the covariance mtx is defined as

$$\Sigma = \begin{bmatrix} \Sigma_{x_1^2} & \Sigma_{x_1x_2} & \Sigma_{x_1x_3} & \cdots & \Sigma_{x_1x_k} \\ \Sigma_{x_2x_1} & \Sigma_{x_2^2} & \Sigma_{x_2x_3} & \cdots & \Sigma_{x_2x_k} \\ \vdots & & & & \vdots \\ \Sigma_{x_kx_1} & \Sigma_{x_kx_2} & \cdots & & \Sigma_{x_k^2} \end{bmatrix}$$

It is a $k \times k$ mtx symmetric mtx

dataset

	X_1	X_2	X_3	...	X_k
1					
2					
⋮					
n					

$$\Sigma = \frac{1}{n-1} \Sigma (x_i - \mu_x)(y_i - \mu_y)^T$$

The principal components are the eigenvectors of the covariance mtx of the data

$x_1 \; x_2 \; x_3 \; x_4 \; x_5$

⋮

n

if dataset of 5 variable

$\Sigma_{5\times5}$ mtx

The first principal component is the eigenvector corresponding to the largest eigenvalue of the covariance mt.

→ If one wants to transform a m-dimensional dataset to a d-dimensional dataset, then one will simply select the first d principle components.

This somewhat remind us of the SVD.

→ The objective of PCA is to perform dimensionality reduction while keeping as much of the information as possible in the high dimensional space.

If we are projecting from M to D dimension PCA will define D vectors, ϕ_D, each of which is N-dimensional.

The d^{th} element of projection x_{nd} (where $x_n = [x_{n1}, \ldots, x_{nD}]^T$) is computed as

$$x_{nd} = \phi_d^T y_n$$

learning task is therefore to choose how many
…nsions one wants to project into D & then pick
projection vector ϕ_D, for each.

PCA uses variance in projected space as the criteria to choose ϕ_D.

$\phi_1 \rightarrow$ be projection that makes variance in x_{n1} as high as possible

Second projected dimension is also chosen to max. variance but ϕ_2 must be orthogonal to ϕ_1

$$\phi_1^T \phi_2 = 0$$

Third component, ϕ_3 again max. variance + orthogonal

$$\therefore \phi_i^T \phi_j = 0 \quad \forall\ i \neq j$$

Further PCA $\rightarrow$ ϕ_i must have a length of 1

$$\phi_i^T \phi_i = 1$$

PCA solves in order to find projection $\phi_1, \ldots,$
and it is derived in no. of ways

A
-l
se
d

$$\bar{y} = \frac{1}{N} \sum_{n=1}^{N} y_n = 0$$

→ force mean to zero.

This is forced by subtracting mean $\bar{y}$ from each y

Finding projection onto $D = 1$ dimen.

We are only interested in finding one ϕ vector

$$x_n = \phi^T y_n$$

Variance $\sigma_x^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{x})^2$ — ①

$$\bar{x} = \frac{1}{N} \sum_{n=1}^{N} \phi^T y_n$$

$$= \phi^T \left(\frac{1}{N} \sum_{n=1}^{N} y_n \right) = \phi^T \bar{y} = 0$$

$\Rightarrow$ ① becomes

$$\sigma_x^2 = \frac{1}{N} \sum_{n=1}^{N} x_n^2$$

Now substituting $x_n = \phi^T y_n$ gives

$$\sigma_x^2 = \frac{1}{N} \sum_{n=1}^{N} (\phi^T y_n)^2$$

$$= \frac{1}{N} \sum_{n=1}^{N} (\phi^T y_n)(y_n^T \phi)$$

$$\sigma_x^2 = \frac{1}{N}\phi^T\left(\frac{1}{N}\sum_{n=1}^{N} y_n y_n^T\right)\phi$$

$$= \phi^T C \phi$$

C is sample covariance mt.

$$C = \frac{1}{N}\sum_{n=1}^{N}(y_n - \bar{y})(y_n - \bar{y})^T$$

but here $\bar{y} = 0$ so.

$$C = \frac{1}{N}\sum_{n=1}^{N} y_n y_n^T$$

As our aim is to find the value ϕ that maximize

σ^2 and ∴ also maximizes $\phi^T C\phi$.

One can say that let keep ↑sing ϕ and max.. $\phi^T C\phi$.

But remember we have constrained ϕ to have

length of 1. $\phi^T\phi = 1$

One can optimize it through the use of Lagrange

multiplier for maximization

condition / constraint of $\phi^T\phi = 1$

$\phi^T\phi - 1 = 0$

$$L = \phi^T C\phi - \lambda(\phi^T\phi - 1)$$

Taking partial derivatives

$$\frac{\partial L}{\partial \phi} = 2C\phi - 2\lambda\phi = 0$$

$C\phi = \cancel{x}\frac{\lambda'}{\cancel{x}}\phi$ $\boxed{\lambda = \lambda'}$

$C\phi = \lambda\phi$

This $= n$ is that of eigenvector

Eigenvalue $= n$,

ϕ is eigenvector of the covariance mt

& λ is eigenvalue.

$\lambda v = A v$

Eigen value $= n$

This suggest us that projection ϕ that maximize the variance is one of the eigenvectors of the Covariance mt C..

However there will be M of these, how do one know which one corresponds to the highest variance?

$\sigma^2 = \phi^T C\phi.$

$\phi^T\phi = 1$ so.

$\sigma^2\phi^T\phi = \phi^T C\phi.$

Removing ϕ^T from both side.

$\sigma^2\phi = C\phi$

$\Rightarrow \sigma^2 \phi = C\phi$

tell us that given an eigenvalue / eigenvector pair (λ, ϕ). λ corresponds to variance of data in the projected space defined by ϕ.

If we find M eigenvector / eigenvalue pairs of the covariance mtx C, the pair with the highest eigenvalue corresponds to the projection with maximal variance ϕ_1. The second highest eigenvalue corresponds to ϕ_2 & so on.

Eigenvectors of the covariance mtx represent the direction (principal components) that capture the most variance in data

Corresponding eigenvalue $\rightarrow$ amount of variance captured by each principal component.

→ Once we have eigenvectors (P.C.); one can transform original data X by projecting it onto new basis formed by eigenvectors

$X_{new} = XV$, V is mtx of eigenvectors

Another way X - be m-dim vector such that

$$X = \sum_{i=1}^{m} y_i \phi_i$$

where $(\phi_1, \phi_2, \ldots, \phi_n)$ form an orthonormal basis of m-dimensional space & the coordinates y_i are given as

$$y_i = \langle X, \phi_i \rangle = X^T \phi_i \quad \forall \ i = 1, 2, \ldots, m$$

Want to represent X with fewer basis vector (let say d where $(d < m)$

One can attempt to do so by replacing $y_{d+1}, \ldots y_m$ with some constant basis

$$\hat{X}(d) = \sum_{i=1}^{d} y_i \phi_i + \sum_{i=d+1}^{m} b_i \phi_i$$

Representation error is

$$\Delta X(d) = X - \hat{X}(d) = \sum_{i=d+1}^{m} (y_i - b_i) \phi_i$$

$$\mathbb{E}(|\Delta X|^2) = \mathbb{E}\left[\left(\sum_{i=d+1}^{m} (y_i - b_i) \phi_i\right)^2\right]$$

$$= \mathbb{E}\left[\sum_{i=d+1}^{m} (y_i - b_i) \phi_i) \sum_{i=d+1}^{m} (y_i - b_i \phi_i]\right]$$

$$= \mathbb{E}\left[\sum_{i=d+1}^{m} \sum_{j=d+1}^{m} (y_i - b_i)(y_j - b_j) \phi_i^T \phi_j\right]$$

as $\phi_i^T \phi_j = 0$ for $i \neq j$

$= 1$ for $i = j$

$$E\left[|\Delta x|^2\right] = \sum_{i=d+1}^{m} E\left[\left[y_i - b_i\right]^2\right]$$

one do $\dfrac{\partial E\left[|\Delta x|^2\right]}{\partial b_i} = 0$

and get $b_i = E[y_i]$

$$E\left[|\Delta x|^2\right] = \sum_{i=d+1}^{m} E\left[y_i - E[y_i]^2\right]$$

$\neq$ $y_i = x^T \phi_i$

$$= \sum_{i=d+1}^{m} E\left[\left(x^T\phi_i - E\left[x^T\phi_i\right]\right)^2\right]$$

$$= \sum_{i=d+1}^{m} \phi_i^T \Sigma_x \phi_i$$

again ~~subject~~ to max. $\dfrac{\partial E}{\partial \phi_i}\left[|\Delta x|^2\right] = 0$ s.t.

$\phi_i^T \phi_i = 1$

one get

$$E\left[|\Delta x|^2\right] = \sum_{i=d+1}^{m} \lambda_i$$

In order to minimize the representation error

λ_i is need to be smallest eigenvalue

In PCA, one choose ~~d~~ eigenvectors corresponding to d largest eigenvalues λ_i of the covariance mtx Σ_x as the principal direction

$$\frac{|\lambda_1| + |\lambda_2| + \dots + |\lambda_d|}{|\lambda_1| + |\lambda_2| + \dots + |\lambda_m|}$$ closer to

1 gives less error.

PCA & SVD

$$\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_p \\ 1 & & & & \\ 2 & & & & \\ \vdots & & & & \\ n & & & & \end{array}$$

ata mt be C of size $n \times p$.

Then principal component are coming from eigenvector of $\Sigma_x = \frac{1}{n-1} C^T C$ is $p \times p$ mt

SVD of C is $\quad C = USV^T$

$$\Sigma = \frac{1}{(n-1)} C^T C = \frac{1}{(n-1)} (USV^T)^T USV^T$$

$$= \frac{1}{(n-1)} VSU^TUSVT = \frac{1}{(n-1)} VS^2V^T.$$

The column of V are the eigenvector of Σ

$\therefore$ If SVD of data mt is $C = USV^T$.

Colun of V gives principal component / direction

PCA

① Calculate covariance mt of data points

② Calculate eigenvector + corresponding eigenvalue.

③ Sort eigenvector in decreasing value.

④ Choose first k eigenvector & that will be new k dimension

⑤ Transform orignl n-dim to k-dimension