Basics

Tuesday, August 25, 2020 7:34 PM

1. Concept:
    - What is PCA?
    - Need of PCA
2. Concept Details:
    - What does it do?
    - Decide on number of components
3. Applications:
    - Reducing the dimension
    - Better visualization
4. Code:
    - Psudo code
    - Links to actual code
5. Pre requisites:
    - Concept of Eigen Value- Eigenvector
6. Resources

Concept

Wednesday, August 26, 2020 12:03 AM

Example: Prediction of the car price

car_ID	symboling	Car Name	fuel type	aspiration	door number	carbody	drive wheel	engine location	wheel base	car length	car width	car height	curb weight	engine type
1	3	alfa-ror	gas	std	two	convertibl	rwd	front	88.6	168.8	64.1	48.8	2548	dohc
2	3	alfa-ror	gas	std	two	convertibl	rwd	front	88.6	168.8	64.1	48.8	2548	dohc
3	1	alfa-ror	gas	std	two	hatchback	rwd	front	94.5	171.2	65.5	52.4	2823	ohcv
4	2	audi 10	gas	std	four	sedan	fwd	front	99.8	176.6	66.2	54.3	2337	ohc
5	2	audi 10	gas	std	four	sedan	4wd	front	99.4	176.6	66.4	54.3	2824	ohc
6	2	audi fox	gas	std	two	sedan	fwd	front	99.8	177.3	66.3	53.1	2507	ohc
7	1	audi 10	gas	std	four	sedan	fwd	front	105.8	192.7	71.4	55.7	2844	ohc
8	1	audi 50	gas	std	four	wagon	fwd	front	105.8	192.7	71.4	55.7	2954	ohc
9	1	audi 40	gas	turbo	four	sedan	fwd	front	105.8	192.7	71.4	55.9	3086	ohc

Difficult to handle:
Memory issues, Multicollinearity, cannot visualize all the variables together

Need of dimension reduction:

- Multiple dimension creates problem on the performance
- Large dataset can lead to system errors
- We cannot visualize multidimensional data

Dimension reduction techniques:

- Low variance filter
- High correlation filter
- Random forest
- PCA

Points to remember:

- PCA is unsupervised technique. But it is used for the supervised algorithms too
- PCA gives the better performance in terms of computation without minor loss information

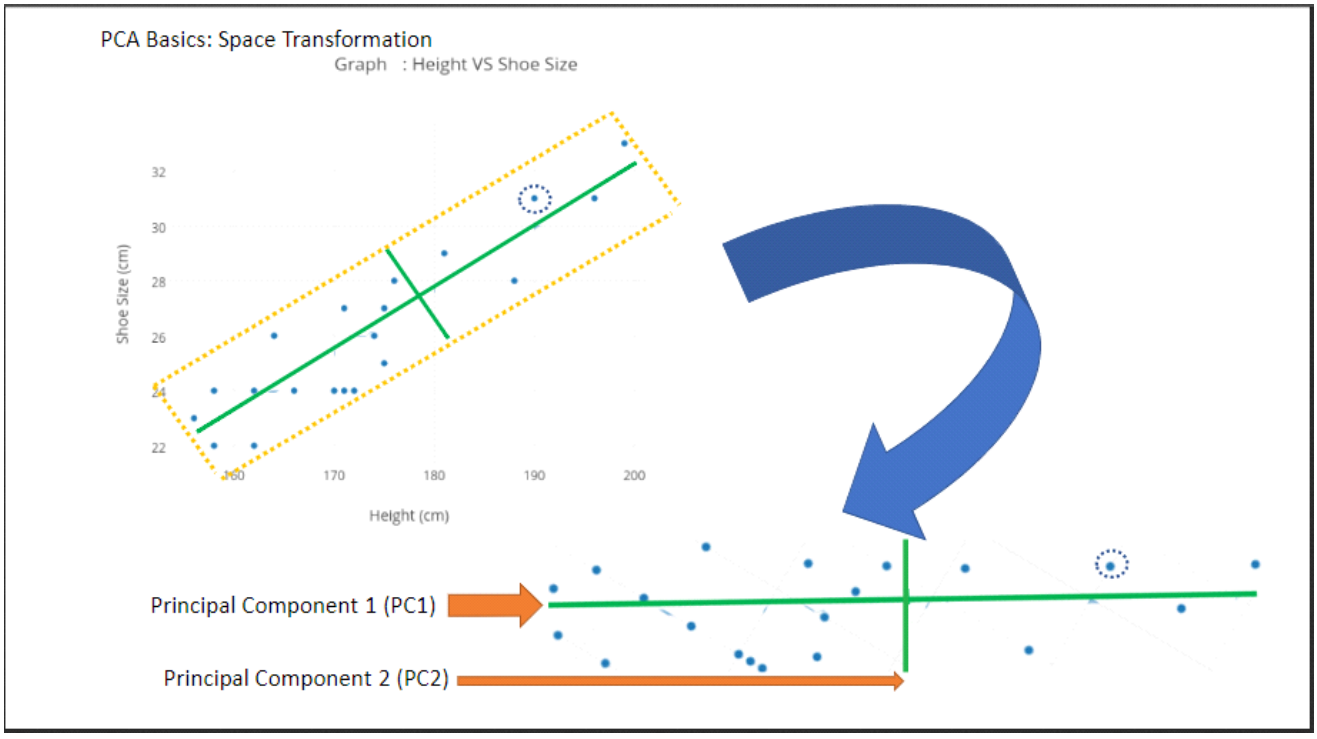
Concept - Details

Home

Wednesday, August 26, 2020 12:03 AM

PCA Uses the concept of the vector rotations and orthogonality of the vectors

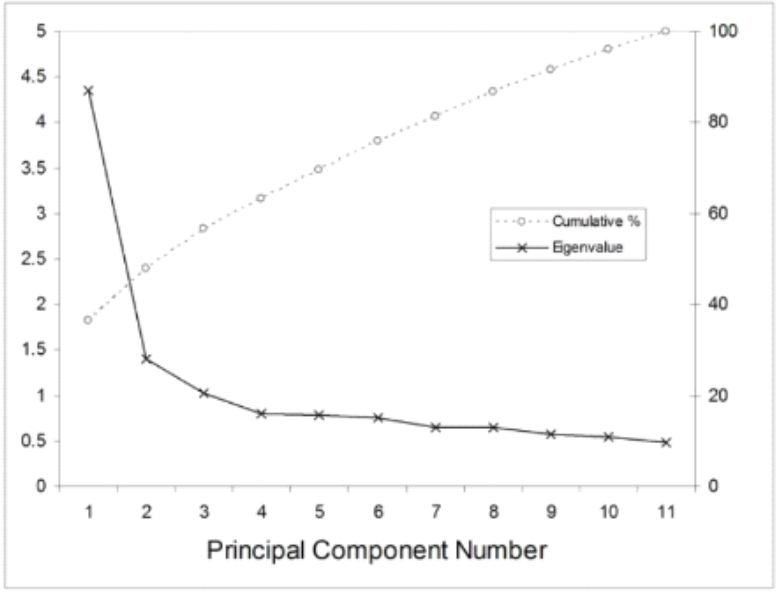
Transformation of the features



What does the PCA do:

How to Decide on how many components to select

1. Usually components are selected so that they can explain ~90% of the variability of the data
2. The components are such that 1st component explains maximum variability and then it goes on decreasing for the rest component s
3. Generally, scree plot is used for deciding the ideal number of components. It is the plot of cumulative variability v/s numbe r of components



What does the Principle components consists of..

1. It is the linear combination of the features
2. The components are orthogonal to each other

Applications

Home

Wednesday, August 26, 2020 12:04 AM

1. Dimensionality reduction
    a. Larger dimension dataset transformation
    b. Produces new features that are “not correlated” can be used for prediction modelling
2. Image related uses
    a. Image compression, reduces size of image
    b. Face recognition (statistical approach for reducing number of variables in face recognition
3. Exploratory Data Analysis
    a. Data visualization

Code

Tuesday, August 25, 2020 7:35 PM

Psudo Code:

1. Standardize the features
2. Compute variance-covariance matrix
3. Find the eigen values and the eigen vector
4. Arrange the eigen values in the descending order
5. Compute the matrix multiplication of the eigen vector with the variance covariance matrix to get the principle components
6. From the scree plot identify number of components sufficient
7. We can train the actual model using these components instead of the original features

Link to the actual codes:

1. Using Sklearn: https://github.com/MadhuraBarve/ML_Models_from_Scratch/blob/master/PCA/03_PCA_Wine.py
2. From scratch: https://github.com/MadhuraBarve/ML_Models_from_Scratch/blob/master/PCA/03_PCA_Wine_Scratch.py

Pre requisites

Wednesday, August 26, 2020 12:04 AM

- Mainly depend on the eigen values and eigen vectors

- If there is a matrix A, then the values of λ satisfying A-λI = 0 Is called as eigen value for the matrix A
- The corresponding equations are the eigen vectors for the matrix A

Resources

Saturday, August 29, 2020 11:56 PM

- https://heartbeat.fritz.ai/understanding-the-mathematics-behind-principal-component-analysis-efd7c9ff0bb3
- youtube.com/watch?feature=player_embedded&v=BfTMmoDFXyE
- https://towardsdatascience.com/the-mathematics-behind-principal-component-analysis-fff2d7f4b643#:~:text=Mathematics%20Behind%20PCA,be%20simplified%20in%20six%20parts%20%3A&text=Compute%20the%20mean%20for%20every%20dimension%20of%20the%20whole%20dataset.