Principal Component Analysis

What is Principal Component Analysis?

- Dimensionality-reduction method, used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

- Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity.

- PCA is simple — reduce the number of variables of a data set, while preserving as much information as possible.

Step by Step Explanation of PCA

- STEP 1: STANDARDIZATION
- STEP 2: COVARIANCE MATRIX COMPUTATION
- STEP 3: COMPUTE THE EIGENVECTORS AND EIGENVALUES OF THE COVARIANCE MATRIX TO IDENTIFY THE PRINCIPAL COMPONENTS

Step 1: Standardization

- Standardize the range of the continuous initial variables so that each one of them contributes equally to the analysis.

- More specifically, the reason why it is critical to perform standardization prior to PCA, is that the latter is quite sensitive regarding the variances of the initial variables. So, transforming the data to comparable scales can prevent this problem.

$$z = \frac{value - mean}{standard\ deviation}$$

Step 2: Covariance Matrix Computation

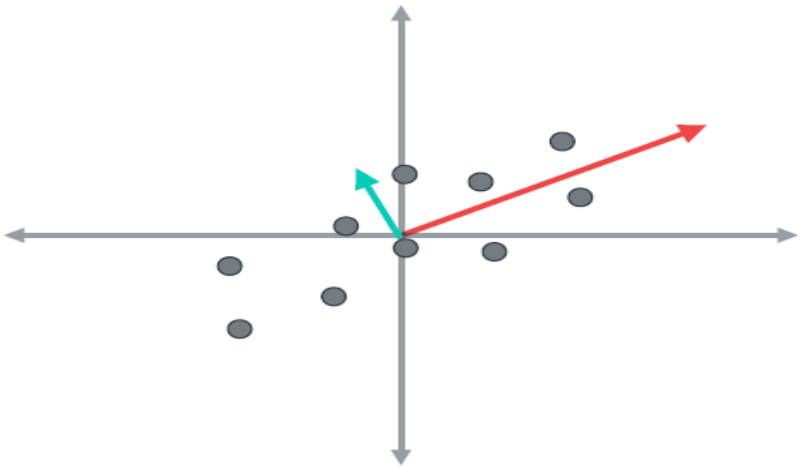
- The aim of this step is to understand how the variables of the input data set are varying from the mean with respect to each other, or in other words, to see if there is any relationship between them.
- Because sometimes, variables are highly correlated in such a way that they contain redundant information.
- So, in order to identify these correlations, we compute the covariance matrix.

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

Step 3: Compute the Eigenvectors and Eigenvalues of the Covariance Matrix

- An ***eigenvector*** of A is a *nonzero* vector v in Rn such that Av=λv, for some scalar λ.
- An ***eigenvalue*** of A is a scalar λ such that the equation Av=λv has a *nontrivial* solution.

Principal Component Analysis (PCA)



$$\Sigma = \begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$$

$$\begin{pmatrix} 2 \\ 1 \end{pmatrix} \quad \begin{pmatrix} -1 \\ 2 \end{pmatrix}$$

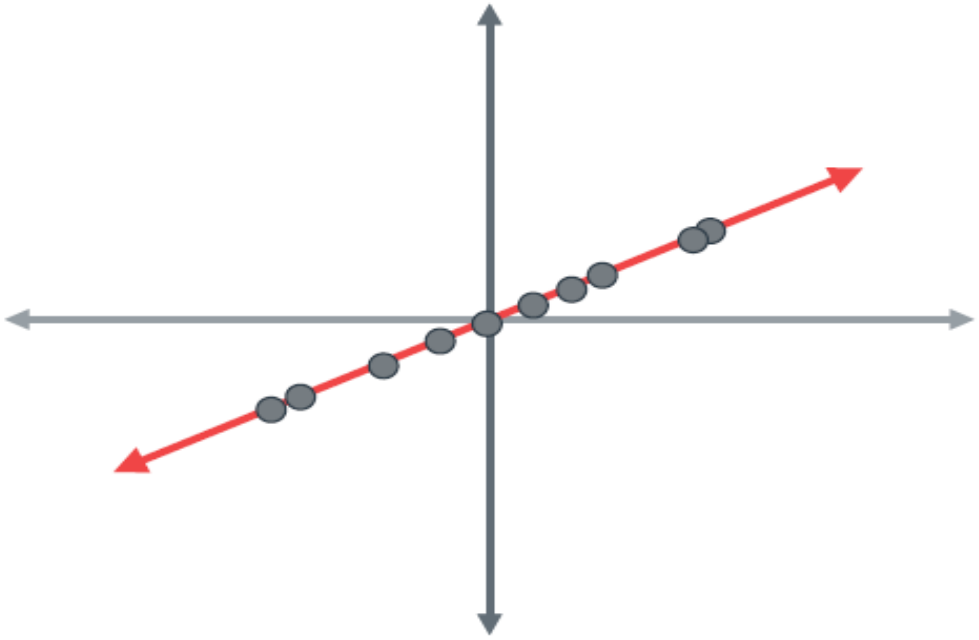
Eigenvectors (direction)

11 1

Eigenvalues (magnitude)

PCA After Reduction

Principal Component Analysis (PCA)



$$\Sigma = \begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$$

$$\begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

Eigenvectors (direction)

$$11$$

Eigenvalues (magnitude)

PCA (During Interviews)

- These are mostly used only for Linear models (Linear Regression and Logistic Regression).
- Stress the importance of Transformation being performed.
- Explain PCA when questions are asked on improving scores of Linear Based Algos and on Feature selection.