Principle Component Analysis (PCA)

→ We use PCA in Preprocessing stage of a model.

→ PCA is one of the process or technique in Dimensionality Reduction.

→Curse of Dimensionality:-

Curse of Dimensionality refers to a set of problems that arise when working with high-dimensional data. The dimension of a dataset corresponds to the number of features that exist in a dataset. A dataset with a large number of attributes, generally of the order of a hundred or more, is referred to as high dimensional data. And the difficulties that come with high dimeonsional data manifest identify patterns while training models and this is called curse of Dimensionality.

→ To overcome the above problem we will use two types of Dimensionality reduction techniques.

1. Feature selection
2. Feature Extraction.

1. Feature Selection:-

In feature selection, the attributes are tested for their worthiness and then selected or eliminated. Some of feature selection technique are:.

1. High correlation filter
2. Feature Ranking
3. Chi Square test.
4. ANOVA test.
5. Backward elimination.

2. Feature Extraction:-

In feature Extraction, the high dimensional attributes are combined in low dimensional components (PCA or ICA) or factored into low dimensional factors (FA).

1. Principal Component Analysis (PCA):-

PCA is a dimensionality-reduction method that is often used to reduce the dimensionality of large datasets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.

→ So, we are reducing no. of variables, while preserving as much information as possible.

steps for performing PCA:-

1. Standardize the range of continuous initial variables.
2. Compute the covariance matrix to identify correlations.
3. Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components.
4. Create a feature vector to decide which principal components to keep.
5. Recast the data along the principal components axes.

Step 1:- STANDARDIZATION

→ The aim of this step is to standardize the range of the continuous initial variables so that each one of them contributes equally to the analysis.

→ More specifically, the reason is if there is large differences between the variables, which will lead to biased results towards larger ranges.

→ Mathematically, this can be done by

$$Z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

→ After standardization, all variables will be transformed to the same scale.

Step2: Covariance Matrix Computation:-

The main aim of this step is to understand how the variables of the input dataset are varying from the mean with respect to each other, or to see if there is any relationship between them.

→ So, inorder to identify the correlations between the variables we compute covariance matrix.

→ The covariance matrix is a pxp symmetric matrix that has entries as the covariances associated with all the possible pairs of the intial variables.

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

p - no. of dimensions

→ Covariance is (+ve) → two variables increase or decrease together (correlated)

→ Covariance is (-ve) → one variable increases when the other decreases (Inversely correlated)

Step3: Compute the Eigen vectors and Eigen values of the covariance matrix to identify Principal components:-

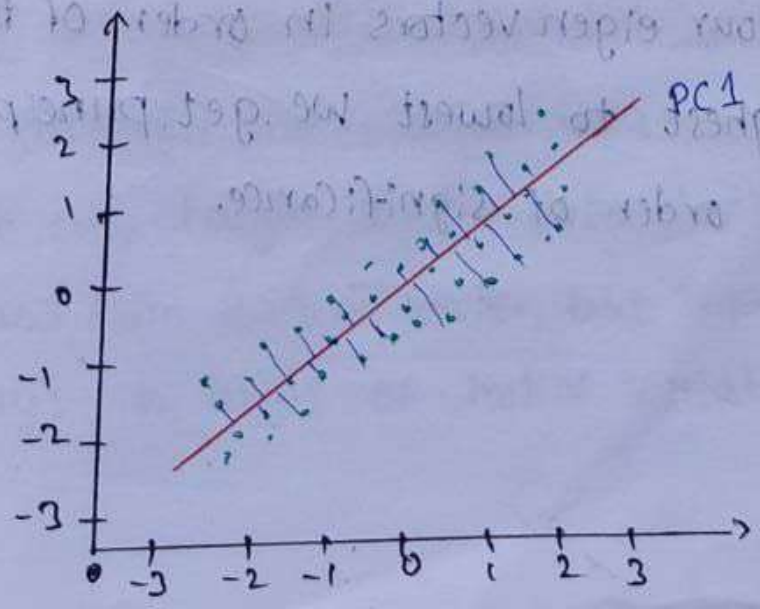
→ Principal components are new variables that are constructed as linear combinations of initial variables.

these combinations are done in such a way that the new variables are uncorrelated and most of the information within the initial variables is squeezed or compressed into first components.

→ So, the idea is 10-dimensional data gives you 10 principal components, but PCA tries to put maximum possible information in the first component,

→ An important thing to realize here is that principal components are less interpretable and don't have any real meaning since they are constructed as linear combinations of the initial variables.

→ PC's are constructed in such a manner that the first principal component accounts for largest possible variance in the dataset. It's the line that maximizes the variance (avg. of squared distances) from the projected points.



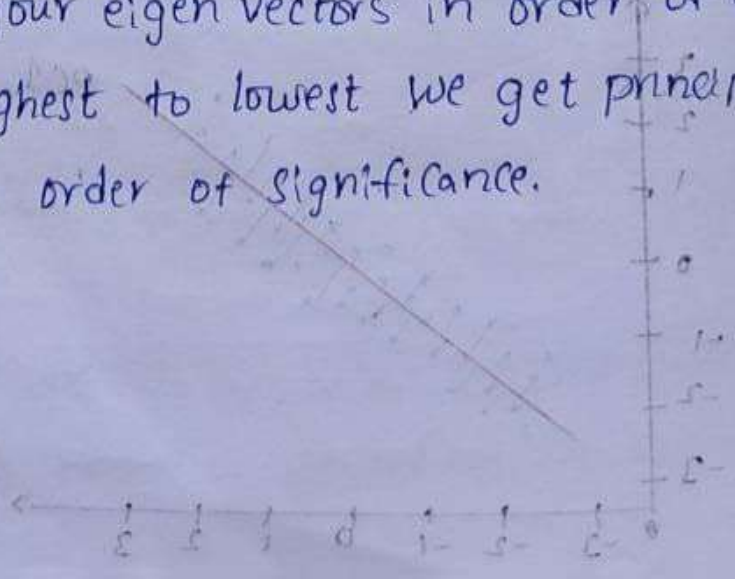
→ The second principal component is calculated in the same way, with the condition that it is uncorrelated (perpendicular to) first component and that it accounts for next highest variance.

→ This continues until total of p→principal components.

→ For every eigen vector has an eigen value. for a 3-dimensional dataset, there are 3 variables, therefore there are 3 eigen vectors with 3 eigen values.

→ The eigen vectors and eigenvalues which are responsible for all the above theory. The eigenvectors of the covariance matrix are actually the directions of the axes where there is most information and that we call Principal Components. And eigen values are simply the coefficients attached to eigen vectors, which give the amount of variance carried in each Principal Component.

→ By ranking your eigen vectors in order of their eigen values, highest to lowest we get principal components in order of significance.



Step4: Feature vector

→ After finding order of significance, In this step, we choose whether to keep all these components or discard those of lesser significance and form with the remaining ones a matrix of vectors that we call Feature vector.

→ so, the feature vector is simply a matrix that has as columns the eigen vectors of the components that we decide to keep. This makes it the 1st step towards dimensionality reduction, because if we choose to keep only p eigen vectors out of n, the final dataset will have only p dimensions.

→ so, Eigen vector v_1 has more information than v_2. ignoring v_2 will have only a loss of exu-4% of the information, the loss will be therefore not important and will still have 96% of information by v_1.

Step5: Recast the data along Principal Components Axes.

→ We just did the standardization and didn't made any changes on the data, just selected PC and form feature vector, but input data remains always in terms of initial variables.

→ In this step, which is the last one, the aim is to use the feature vector formed using eigenvectors of the covariance matrix, to orient the data from the original axes to the ones represented by principal components (hence name PCA). This can be done by multiplying the transpose of original dataset by the transpose of feature vector.

Final Dataset = Feature Vector^T × Standardized original Dataset^T