Principal components Analysis developed by Karl pearson in 1901.

PCA combines the attributes and hence reducing the dimension of a data set.

The basic procedure is as follows.

1. Load the data set 'A'
2. Normalized the data. Generally zero-mean normalization is done. $DataAdjust = \frac{\sum_{i=1}^{n} x_i - \bar{x}}{\sigma}$
3. Find the covariance of the matrix A.
   $COV(X,Y) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})/n-1$
4. Find the Eigen vector (U) and Eigen value (λ) of the covariance matrix A.
   The eigen vector with highest eigen value is the principal component of the data.
5. Arrange the eigen value in decreasing order.
6. Arrange the eigen vector according to the eigen value.
7. Keep only these eigen vector for which the eigen value is > 1.
8. Multiply this reduced eigen vector matrix with the normalized data set, it gives the principal components of the data sets.

Factor Analysis

1. Load the data set
2. Obtain the normalize data set (Y) by substracting the column wise mean from each sample in the column and divide it by the column wise standard deviation.
3. calculate the correlation matrix (R) using the relation

$$R = \frac{Y \times Y'}{n-1}, \quad n = \text{total no. of samples.}$$

4. calculate the eigen vector (U) and Eigen value (λ) from R.
5. Rearrange the Eigen values and corresponding Eigen vector in descending order.
6. calculate the factor loading matrix (A) using $A = U \times \sqrt{\lambda}$
7. calculate the factor score (F)

$$B = R^{-1} \times A$$

$$\boxed{F = B \times Y.}$$

B = Factor coefficient matrix.

R^{-1} = inverse of the correlation matrix.

A = Factor loading matrix.

Curse of dimensionality

→ When dimensionality increases, data become increasingly sparse → Information content
→ Density and distance between points, w
is critical to clustering, outlier analy
becomes less meaningful.
→ The possible combinations of subspaces
grow exponentially.

Dimensionality Reduction

→ Avoid the curse of dimensionality.
→ Help eliminate irrelevant features
and reduce noise.
→ Reduce time and space required i
data mining.
→ Allow easier visualization.

Dimensionality reduction techniques

DFT
DCT
DWT
PCA
ICA.

Lossless - zip files
Compression image format. gif, tiff and png