ASSIGMENT 6

Aim: Principal component Analysis- Finding Principal Components, Variane and standard Deviation calculations of principal components

Theory:

Principal component analysis is a method of extracting important variables from a large set of variables available in a data set. It extracts low ~~di~~ dimensional set of features from a high dimensional data set with a motive to captre as much information as possible. With fewr variables, visualisation becomes much more meaningful. PCA is more useful when dealing with 3 or higher dimensional data.

It is always performed on a symmetric correlation or covariance matrix. This means the matrix should be numeric and have standardized data.

The principal components are supplied with normalized version of original predictors. This is because the original predictors may have different scales For example, imagine a data set with variables measuring units as gallons, kilometres, light years, etc. It is definite that the scale of variances in these variables will be large.

Performing PCA on un-normalized variables will lead to insanely large loading for variables with high variance. In turn, this will lead to dependence of a principal component on the variable with high variance.

Highlights:

1. PCA is used to overcome features redundancy in a data set
2. These features are low dimensional in nature.
3. These features ~~the~~ components are a resultant of normalized linear combination of original predictor values.
4. These components aim to capture as much information as possible with high explained variance.
5. The first component has the highest variance followed by second, third and soon.
6. The components must be uncorrelated
7. Normalizing data becomes extremely important when the predictors are measured in different units.
8. PCA is applied on a data set with numeric variables.

Conclusion: PCA using R was implemented in this assignment successfully.