WEEK-3 REPORT P-1

DASARI SRINITH (CS21BTECH11015)

PCA:

Theory:

- Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often 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.
- The idea of PCA is simple — reduce the number of variables of a data set, while preserving as much information as possible.
- There are mainly 5 steps in PCA
    1. Standardization : This can be done by subtracting the mean and dividing by the standard deviation for each value of each variable.
    2. Covariance Matrix Computation : The covariance matrix is a p $\times$ p symmetric matrix (where p is the number of dimensions) that has as entries the covariances associated with all possible pairs of the initial variables.
    3. Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components : Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables.These combinations are done in such a way that the new variables (i.e., principal components) are uncorrelated and most of the information within the initial variables is squeezed or compressed into the first components.
    4. Feature Vector : The feature vector is simply a matrix that has as columns the eigenvectors of the components that we decide to keep.
    5. Recast the datat along the principal components : This can be done by multiplying the transpose of the original data set by the transpose of the feature vector
- As there are as many principal components as there are variables in the data, principal components are constructed in such a manner that the first principal component accounts for the largest possible variance in the data set.
- The second principal component is calculated in the same way, with the condition that it is uncorrelated with (i.e., perpendicular to) the first principal component and that it accounts for the next highest variance.

1. Assumptions taken in PCA.

   Solution:

   (a) There needs to be a linear relationship between all variables.

   (b) Your data should be suitable for data reduction.

   (c) You should have sampling adequacy

2. Limitations of PCA.

   Solution:

   (a) Doesn't work well for non linearly correlated data.

   (b) Always finds orthogonal principal components. Sometimes, our data demands non-orthogonal principal components to represent the data.

   (c) If the variables are correlated, PCA can achieve dimension reduction. If not, PCA just orders them according to their variances

3. Is rotation necessary in PCA?

   Solution:
   Yes

4. Can we use PCA for feature selection?

   Solution:
   In PCA, we obtain Principal Components axis, this is a linear combination of all the original set of feature variables which defines a new set of axes that explain most of the variations in the data.It does not result in the development of a model that relies upon a small set of the original features and so for this reason, PCA is not a feature selection technique.

5. What's the difference between PCA and t-SNE?

   Solution:
   PCA , tries to preserve global structure/Shape ; but t-SNE tries to preserve the local structure by taking into account the distance between the point. PCA , is a linear technique: reduce the dimension of the data when the linear correlations are strong ; but t-SNE is a non-linear technique so it can interpret the complex polynomial relationships between features.