

UROP Report 1

(Sept)

# **PCA Algorithm**

Principal Component Analysis (PCA) is one of the most commonly used un-supervised machine learning algorithms across a variety of applications like : exploratory data analysis, dimensionality reduction, information compression and more. Reducing the number of variables of a data set naturally comes at the expense of accuracy. Because smaller data sets are easier to explore and visualize and make analyzing data much easier and faster for machine learning algorithms without external variables to process. The main idea of PCA is to reduce the number of variables of a data set, while preserving more information as much as possible.

Step by step explanation of PCA:

Step-1:- First we have to do standardize the range of the continuous initial variables so that each one of them contributes equally to the analysis. Mathematically, this can be done by subtracting the mean and dividing by the standard deviation for each value of each variable or by calculating the mean for the data sets.

Step-2:- After the standardization is done successfully. Now, all the values/variables will be transformed to the same scale. In this step we will do covariance matrix computation. Why we want to compute? What is the use of these computation? Let, discuss how it helps. 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. The covariance matrix is a m \* m symmetric matrix( where m is the number of dimensions). The number of covariance values are 2 power n(2^n). since the covariance of a variable with itself is its variance (cov(a,a)=var(a)) and so on. W.K.T the covariance is not more than a table that summarizes the correlations between all the possible pairs of variables.

Step-3:- In this step we compute the eigen vectors and eigen values of the covariance matrix to identify the principal components. Eigen vectors and eigen values are the linear algebra concepts based. Here, the principal components are new variables that are constructed as linear combination or mixtures of the initial variables. These combinations are done in such a way that the new variables are uncorrelated and mist of the information within the initial variables is squeezed or compressed into the first components. PCA tries to put maximum possible information in the first component. Let see about why we want know about eigen vectors and eigen values. In this concept we know that every eigen vector has an eigen value and their number is equal to the number of dimensions of the data. For example, we have a 2-dimensional data set and in this data set we can see 2 variables, 3 eigen vectors with corresponding 2 eigen values. By ranking your eigen vectors in order of their eigen values from highest to lowest we get the principal components in order of significance. For example if we got 2nd eigen value is greater than 1st eigen value then we take the big value for next step calculation.

Step-4:- In this step we will compute the values into normalized vectors or feature vectors. By computing the normalized vectors we get a n\*m matrix( here m=1) with the normalized vector values we form a equation with the help of variables. i.e., x(position 1) + y(position 2)+….. here, x, y,… values are normalized vector values.

The main idea we understand that PCA simplifies the complexity in high-dimensional data while retaining treads and patterns. It improves the performance of ML algorithm as it eliminates correlated variables that don’t contribute in any decision making.