Implementation of Principal Component Analysis

Lab Assignment 7

Sumedha Janani Siriyapuraju
Dept. of Electronics and Communication Engineering
Visvesvaraya National Institute of Technology
Nagpur, India
sumedhasjs@gmail.com

***Abstract*—Principal component analysis (PCA) is a powerful mathematical technique to reduce the complexity of data. It detects linear combinations of the input fields that can best capture the variance in the entire set of fields, where the components are orthogonal to and not correlated with each other. The goal is to find a small number of derived fields (principal components) that effectively summarize the information in the original set of input fields.**

***Index Terms*—Principle Component Analysis,**

I. Introduction

PCA is an advanced algorithm for data exploration that you can use to find patterns in the data and to identify a transformed representation of data that highlights these patterns.

PCA is based on an orthogonal, linear transformation of data into a new representation space. It can be thought of as a replacement of the original attributes by new attributes, so-called principal components. These principal components correspond to the directions in the original attribute space that shows the greatest variance.

The number of principal components that are used is at most identical to the number of original attributes. Often, however, the number is considerably lower, because one goal of PCA is the reduction of dimensionality. Whereas simple algorithms for data exploration might be sufficient for exploring single attributes or pairs of attributes, PCA is most useful for multidimensional data that has several attributes. For multidimensional data, simple algorithms are not sufficient.

If the analyzed data set shows strong patterns, the dimension of the new representation that is obtained by using PCA can be considerably reduced without significant loss of information.

II. Method

- Standardize the range of continuous initial variables. This can be done by subtracting the mean and dividing by the standard deviation for each value of each variable.

$$z = \frac{value - mean}{standard\ deviation}$$

- Compute the covariance matrix to identify correlations. The covariance matrix is a $p*p$ symmetric matrix (where p is the number of dimensions) that has as entries the covariances associated with all possible pairs of the initial variables.
- Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components
- Create a feature vector to decide which principal components to keep Recast the data along the principal components axes.In this step, what we do is, to choose whether to keep all these components or discard those of lesser significance (of low eigenvalues), 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 eigenvectors of the components that we decide to keep.
- Recast the data along the principal component axes.

$$final\ dataset = feature\ vector^{\ T} * Standardized\ Dataset^{\ T}$$

III. Discussion

- Principal component analysis, or PCA, is a statistical procedure that allows you to summarize the information content in large data tables by means of a smaller set of “summary indices” that can be more easily visualized and analyzed.
- PCA forms the basis of multivariate data analysis based on projection methods.
- The most important use of PCA is to represent a multivariate data table as smaller set of variables (summary indices) in order to observe trends, jumps, clusters and outliers.

IV. Conclusion

A function for performing the operations in Principal Component Analysis has been successfully implemented.

Appendix (Code)

the code is available here