IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

6th National Conference on Science, Technology and Communication Skills – NCSTCS 2K23

Narula Institute of Technology, Agarpara, Kolkata, India

Vol. 10, Special Issue 3, September 2023

Data Mining by Dimension Reduction Using Principal Component Analysis

Shilpi Pal[1], Apurba Ghosh[1], Gargi Bose[2] and Rajarshi Nath[2]

Faculty, Department of Basic Science and Humanities, Narula Institute of Technology, Kolkata, India[1]

Student, Department of Computer Science and Engineering, Narula Institute of Technology, Kolkata, India[2]

Abstract: This paper provides a complete and simplified explanation of how to reduce the dimension (attributes) of large data, keeping the data's importance intact. Thus in this paper, we have considered a modern technique of dimension reduction which is widely known as principal component analysis (PCA). Here we have observed how PCA works step by step, by dividing the whole process number of simple steps. We have gone through each step, providing a logical explanation of what PCA is doing and simplifying mathematical concepts such as standardization, covariance, Eigenvalue and Eigenvectors along with focusing on how the compute them. PCA is a widely used in data mining technique for extracting information from large datasets which is done by reducing the dimension of data to extract the important features and reduce the complexity of the data. So we have taken a numerical example to illustrate the model easily.

Keywords: Principal component analysis, Dimension reduction, Data mining.

I. INTRODUCTION

Dimension reduction [1] is a necessary step in the effective analysis of massive high-dimensional data sets. It may be the main objective in the analysis for visualization of the high-dimensional data or it may be an intermediate step that enables some other analysis such as clustering. Principal component analysis [2] (PCA) was first introduced by Pearson [3] in 1901 and later independently developed by Tuffery [4] in 1933, where the name principal components first appears. PCA is probably the oldest and certainly the most popular technique for computing lower-dimensional representations of multivariate data. The technique is linear in the sense that the components are linear combinations of the original variables (features), but non-linearity in the data is preserved for effective visualization. The technique can be presented as an iterative computation of the direction of the highest variation followed by projection onto the perpendicular plane. This quickly provides a few perpendicular directions that account for the majority of the variation in the data, giving a low-dimensional representation of the data.

Dimensionality reduction and data mining using statistics [5] have evolved significantly. In the 1930s, Pearson [3] introduced Principal Component Analysis (PCA) for reducing data dimensions. In the 1960s, Feigenbaum [6] and others laid the groundwork for knowledge representation in data mining. Dimensionality reduction progressed with Isomap by Tenenbaum et al. [7] in the 2000s. These scientists' contributions have been instrumental in shaping the fields of dimensionality reduction and Data Mining [8]. 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. Reducing the number of variables of a data set naturally comes at the expense of accuracy, but the trick in dimensionality reduction is to trade a little accuracy for simplicity. Because smaller data sets are easier to explore and visualize and make analyzing data much easier and faster for machine learning algorithms without extraneous variables to process. PCA is used to identify the correlation between variables, detect outliers and anomalies and generate new features from the existing ones. PCA can also be used in the process of feature selection whereby, the most relevant are selected for further analysis. Additionally, PCA can be used in the process of pattern recognition, where patterns in a dataset can be identified and used to make predictions or generate insights. So, to sum up, the idea of PCA is simple — reduce the number of variables of a data set, while preserving as much information as possible. The Data Mining [3] process using Principal Component Analysis (PCA) can involve several steps.

In this paper, first, the data is pre-processed, including feature selection and normalization, to prepare it for the PCA algorithm. Next, the PCA algorithm is applied to the data to identify the dimensions of greatest variance in the data. Then, the data is reduced to the identified dimensions and used to create a model. Finally, the model can be evaluated and used to make predictions. In paper writing, explaining how PCA works and how it can be used to analyze and interpret large datasets is important. PCA is a statistical technique that captures the correlation among multiple variables and expresses the data in a form that is easier to interpret. It is a powerful tool for data visualization and for identifying patterns and trends in large datasets. PCA can be used for finding correlations, for feature selection, for dimensionality reduction, and for feature extraction. Additionally, PCA is useful for data compression and for reducing the complexity of large datasets

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

6th National Conference on Science, Technology and Communication Skills – NCSTCS 2K23

Narula Institute of Technology, Agarpara, Kolkata, India

Vol. 10, Special Issue 3, September 2023

Reducing the dimension by Principal Component Analysis (PCA) is a popular linear dimension reduction algorithm used in machine learning and data analysis to reduce the dimensionality of a dataset. It works by projecting high-dimensional data into a lower-dimensional space while preserving the variance in the data.

PCA tries to find a new set of features (principal components) that explain most of the variance in the original data set. These principal components are linear combinations of the original features and are orthogonal to each other. By keeping only the top n principal components, PCA can significantly reduce the dimension of a dataset. This can be useful for visualization and reducing the computational complexity of a machine learning model.

II. PRELIMINARIES

Arithmetic mean: The arithmetic mean (also known as the mean or average) is a statistical measure of central tendency, which represents the typical or central value of a dataset. It is calculated by adding up all the values in a dataset and then dividing the sum by the number of values in the dataset. The arithmetic mean is widely used in a variety of fields such as mathematics, statistics, economics, and science to understand and analyze data.

For example, if you have a dataset of 10 values, say [2, 4, 6, 8, 10, 12, 14, 16, 18, 20], the arithmetic mean can be calculated by adding up all the values and dividing the sum by 10. (2+4+6+8+10+12+14+16+18+20)/10 = 11

In this case, the arithmetic mean is 11. This implies that in this dataset, a typical value can be approximated by 11. It's worth mentioning that the arithmetic mean is more sensitive to outliers than other measures of central tendency, such as the median or mode. The arithmetic mean is commonly used in Principal Component Analysis (PCA) to calculate the mean of the data prior to conducting the analysis. In PCA, the mean is subtracted from each data point so that the mean of the resulting data is zero. This is done to center the data about the origin, which helps to maximize the variance explained by the principal components. Specifically, in PCA, the arithmetic mean of each variable is subtracted from every observation in that variable. The resulting data is then used to calculate the covariance matrix, which is the basis for the principal components.

It's worth noting that other types of means, such as the geometric mean or the harmonic mean, can also be used in PCA. However, the arithmetic mean is the most commonly used mean in this context.

Covariance matrix: In probability theory and statistics, a covariance matrix is a square matrix that captures the covariance between each pair of elements in a given random vector. More specifically, if you have a random vector X with n elements, the covariance matrix C of X is an [n x n] matrix where the (i, j)th element of C represents the covariance between the ith and jth elements of X. The diagonal elements of the covariance matrix represent the variances of the individual elements of X.

The formulae of the Co-variance matrix are

$$Cov(x,x) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - x)^2 \text{----------(1)}$$

$$Cov(x,y) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - x)(y_i - y) \text{----------(2)}$$

$$Cov(y,y) = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - y)^2 \text{----------(3)}$$

The covariance matrix is an important concept in statistics and machine learning, as it is used to represent the relationships between different variables and to understand the spread and correlation of the data. For example, in Principal Component Analysis (PCA), the covariance matrix is calculated to determine the principal components of the data.

The covariance matrix is also used in linear regression to calculate the coefficients of the model and to determine the extent to which the independent variables are related to the dependent variable.

It's worth noting that the covariance matrix can be affected by outliers, and that it may be necessary to remove them from the data or to use robust methods to estimate the matrix in some cases.

Eigen value and its Use in PCA: The eigenvalues and eigenvectors play a fundamental role in Principal Component Analysis (PCA). The eigenvectors and eigenvalues of the covariance matrix are used in PCA to identify the principal components of a dataset. A principal component is a linear combination of the original variables and represents a direction in which the data varies the most. The first principal component captures the maximum amount of variation in the original data, and each subsequent principal component captures as much of the remaining variation as possible in a mutually orthogonal direction.

The eigenvectors of the covariance matrix represent the directions in which the data has the most variation, and the eigenvalues associated with each eigenvector represent the magnitude of the variation along that direction. Thus, the eigenvectors with the highest eigenvalues are considered the most important principal components.

PCA involves calculating the eigenvectors and eigenvalues of the covariance matrix and then sorting them in descending order of eigenvalues. These eigenvectors are chosen as the principal components and can be used to reduce the dimensionality of the dataset. PCA is a powerful tool in machine learning and data analysis as it reduces the dimensionality of high-dimensional data and helps identify underlying patterns or structures in the data.

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

6th National Conference on Science, Technology and Communication Skills – NCSTCS 2K23

Narula Institute of Technology, Agarpara, Kolkata, India

Vol. 10, Special Issue 3, September 2023

Eigen vector and its Use in PCA: Normalized eigenvectors are used in PCA to ensure that the principal components have unit lengths, which makes them easier to interpret and makes the results more stable.

In PCA, the eigenvectors of the covariance matrix represent the directions in which the data has the most variation. However, the length of the eigenvectors is not necessarily unity, and this can lead to numerical instability in some situations. Therefore, it is common practice to normalize the eigenvectors to have a unit length. After computing the eigenvectors and eigenvalues of the covariance matrix, the eigenvectors can be sorted in order of descending eigenvalues. Then, each eigenvector is normalized to have a unit length. The resulting normalized eigenvectors represent the principal components of the data.

Using normalized eigenvectors ensures that each principal component has a unit length and thus has the same level of influence in the analysis. It also improves the numerical stability of the analysis.

In summary, the normalization of eigenvectors is a key step in PCA because it ensures that a Principal Component (PC) is Strictly *Orthogonal* and increases its stability, comparability of the components by assigning them all equal influence, and interpretability.

III. ALGORITHM FOR DATA REDUCTION USING PCA

An algorithm is a step-by-step procedure for solving a problem or performing a task. It enhances efficiency, accuracy, and consistency in various fields such as computing, data analysis, and problem-solving, enabling streamlined processes and optimal outcomes.

Step 1: Start the program.

Step 2: Prompt the user to enter the number of rows and columns of the matrix.

Step 3: Create afor loop to request the user to enter each element of the matrix and store it in a 2D array.

Step 4: Compute and display the covariance matrix of the transpose of the input matrix.

Step 5: Compute the eigenvalues and eigenvectors of the covariance matrix and save them to variables.

Step 6: Compute the average of each row of the input matrix.

Step 7: Subtract the average from each element in each row of the input matrix and compute the resulting matrix.

Step 8: Compute all principal components of the covariance matrix.

Step 9: Extract only the desired number of principal components.

Step 10: Display the computed principal components.

Step 11: End the program.

IV. NUMERICAL EXAMPLE

From the above discussion we already know we reduce data by using PCA.Now we have 2 data set and each data set has four element. We try to make it one data set of four elements.

Use PCA Algorithm for the following data to reduce the dimension from 2 to 1

Feature	E-1	E-2	E-3	E-4
X	4	8	13	7
Y	11	4	5	14

Now reducing the given data set using PCA:
Here the number of samples N=4 and the datasets(n)=2;
X = (4+8+13+7)/4=8
Similarly, Y = 8.5
Covariancematrix:-

$$[cov(x,x)\ cov(x,y)\ cov(y,x)\ cov(y,y)\ ]$$

Recalling the equation (1) from the preliminaries of Covariance-matrix
Cov (x, x) = 14
Recalling the equation (1) from the preliminaries of Covariance-matrix
Cov (x, x) = -11
Recalling the equation (1) from the preliminaries of Covariance-matrix
Cov (y, y) =23

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

6th National Conference on Science, Technology and Communication Skills – NCSTCS 2K23

Narula Institute of Technology, Agarpara, Kolkata, India

Vol. 10, Special Issue 3, September 2023

Thus, the covariance matrix is:

$$A = [14 \ -11 \ -11 \ 23 \ ]$$

Now we calculate the eigenvalue of the corresponding covariant matrix A

Her equation $|A - \lambda I|$=0------(A)

$$\left|14 - \lambda \ -11 \ -11 \ 23 - \lambda \right| = 0$$

For the largest Eigen Value 30, now we find the eigenvector by using the following equation,
$(14- \lambda)\ x_1-11x_2=0$
$-11x_1+(23-\lambda)\ x_2=0$ (A)
Solving A
$x_1=11t$, $x_2=(14-\lambda)\ t$
So corresponding to the largest eigenvalue, the eigenvector is: x1=11, x2=-16.3849
i.e., $X = [11 \ -16.3849 \ ]$
Thus normalized vector corresponds to the largest eigenvalue is denoted as e_1 and it is given by e_1=$[0.5574 \ -0.8303 \ ]$
Now derive the new dataset to obtain the first principal component we have to compute:

$$P_{1i} = e_1^T \left[x_i - \underline{x} \ y_i - \underline{y} \right]$$

Thus,
P_{11} =$e_1^T[4 - 8 \ 11 - 8.5 \ ]$= [0.5574 -0.8303]$[-4 \ 2.5 \ ]$=-4.3053
P_{12} =$e_1^T[8 - 8 \ 4 - 8.5 \ ]$= [0.5574 -0.8303]$[0 \ -4.5 \ ]$=-3.7363

P_{13} =$e_1^T[13 - 8 \ 5 - 8.5 \ ]$= [0.5574 -0.8303]$[5 \ -3.5 \ ]$= 5.693

P_{14} =$e_1^T[7 - 8 \ 14 - 8.5 \ ]$= [0.5574 -0.8303] $[-1 \ 5.5 \ ]$= -5.123

First PC	-4.305	3.736	5.693	-5.124

V. CONCLUSION

Using PCA, we can identify the most important factors or features in a dataset, reduce the dimensionality of the dataset while retaining as much of the original variation as possible, and identify underlying patterns or structures in the data. This makes PCA a powerful tool for data mining, exploratory data analysis, and feature selection. By using PCA, we can draw conclusions about the most important factors or features in a dataset, which can help guide further analysis or decision-making. PCA is used to identify the correlation between variables, detect outliers and anomalies and generate new features from the existing ones. PCA can also be used in the process of feature selection whereby, the most relevant are selected for further analysis. Additionally, PCA can be used in pattern recognition, where patterns in a dataset can be identified and used to make predictions or generate insights. **Further scopes:** Further research directions in data mining by PCA could include exploring ways to improve the interpretability and the explainability of PCA results, as well as developing novel machine learning algorithms that incorporate PCA as a preprocessing step. Additionally, there may be opportunities to apply PCA to new types of data or to investigate the effectiveness of PCA in different domains or applications. Another potential avenue for research is to investigate the impact of different normalization and scaling techniques on PCA results. Finally, further investigation into the relationship between the number of principal components used and the overall performance and interpretability of PCA results can also be pursued.

REFERENCES

[1]. Han J., Kamber M. and Pei J. (2006). Dimension Reduction: Concepts and Techniques. Morgan Kaufmann is an imprint of Elsevier, USA.
[2]. Jolliffe I.T. (1973). Principal Component Analysis. Journal of the Royal Statistical Society Series, New York, Berlin, Heidelberg: Springer-Verlag, Second Edition.
[3]. Pearson K. (1901). Lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11), 559-572.
[4]. Tufféry S. (2011). Data Mining and Statistics for Decision Making. Wiley Publication, ISBN: 9780470688298.
[5]. Siebes A. (2000). Data Mining and Statistics. In: Della Riccia, G., Kruse, R., Lenz, HJ. (eds) Computational Intelligence in Data Mining. International Centre for Mechanical Sciences, 408, 1-38.
[6]. Feigenbaum E. (1981). The Handbook of Artificial Intelligence. First edition, Stanford University, Heuris Tech Press and William Kaufmann, Inc., California.
[7]. Tenenbaum J., Kemp C., Griffiths T.L. and Goodman N. D. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. Science, 331(6022), 1279-85.
[8]. Hand D.J., Mannila H., and Smyth P. (2007). Principles of Data Mining. Drug safety, 30, 621-622.