UNIVERSIDAD POLITÉCNICA DE YUCATAN

ROBOTICS COMPUTATIONAL ENGINEERING

MACHINE LEARNING

PROFESSOR: VICTOR ALEJANDRO ORTIZ SANTIAGO

STUDENTS:

FREDY EDUARDO ALONZO MONDRAGON

JESUS GABRIEL CANUL CAAMAL

JESUS EDUARDO CASAS NAVARRO

ESTEBAN RODRIGUEZ CUMUL

GRADE: 9° GROUP: B

SCHOOLAR CICLE: 2023-2024 B

PRINCIPAL COMPONENT ANALYSIS

Dimensionality Reduction:

In machine learning we are having too many factors on which the final classification is done. These factors are basically, known as variables. The higher the number of features, the harder it gets to visualize the training set and then work on it. Sometimes, most of these features are correlated, and hence redundant. This is where dimensionality reduction algorithms come into play.

Components/Factor Based:

Feature selection:

In this, we need to find a subset of the original set of variables. Also, need a subset which we use to model the problem. It usually involves three ways:

Filter, Wrapper, Embedded

Feature extraction:

We use this, to reduces the data in a high dimensional space to a lower dimension space, i.e. a space with lesser no. of dimensions.

Application of PCA in machine learning:

- Dimensionality Reduction: Simplifying high-dimensional data.
- Noise Reduction: Filtering out noise and redundancy.
- Feature Engineering: Creating new informative features.
- Visualization: Projecting data for understanding and clustering.
- Preprocessing: Standardizing data and handling multicollinearity.
- Compression: Reducing data size while preserving information.
- Face Recognition: Reducing image dimensionality for facial analysis.
- Anomaly Detection: Identifying unusual data points.
- Collaborative Filtering: Enhancing recommendation systems.
- Speech Recognition: Reducing dimensionality of audio features.

Principal Components analysis method:

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 points much easier and faster for machine learning algorithms without extraneous variables to process.

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.

Step-by-Step Explanation of PCA

Step 1: Standarization

The aim of this step is to standardize the range of the continuous initial variables so that each one of them contributes equally to the analysis.

More specifically, the reason why it is critical to perform standardization prior to PCA, is that the latter is quite sensitive regarding the variances of the initial variables. That is, if there are large differences between the ranges of initial variables, those variables with larger ranges will dominate over those with small ranges (for example, a variable that ranges between 0 and 100 will dominate over a variable that ranges between 0 and 1), which will lead to biased results. So, transforming the data to comparable scales can prevent this problem.

Step 2: Covariance Matrix Calculation

The aim of this step is to understand how the variables of the input data set are varying from the mean with respect to each other, or in other words, to see if there is any relationship between them. Because 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.

Calculate the covariance matrix of the standardized data. The covariance matrix measures the degree to which two variables change together. For a dataset with p features, the covariance matrix will be a p x p square matrix.

Step 3: Eigendecomposition of the Covariance Matrix

Perform eigendecomposition (also known as eigenvalue decomposition) on the covariance matrix. This step yields a set of eigenvalues and their corresponding eigenvectors.

Each eigenvalue represents the variance explained by its corresponding eigenvector.

Eigenvectors are the directions in which the data varies the most. These eigenvectors are the principal components.

Step 4: Selection of Principal Components

Sort the eigenvalues in descending order. The eigenvector corresponding to the highest eigenvalue explains the most variance in the data and becomes the first principal component.

Continue selecting eigenvectors in descending order of their eigenvalues to form a sequence of principal components.

Step 5: Dimension Reduction

Choose the number of principal components you want to retain. This is often based on the proportion of total variance you want to preserve. A common choice is to retain enough principal components to explain, for example, 95% or 99% of the total variance.

These selected principal components will form a new coordinate system for your data.

Step 6: Projection

Project the original data onto the new coordinate system formed by the selected principal components. This involves multiplying the original data matrix by the matrix of selected eigenvectors.

Step 7: Interpretation and Analysis

Analyze the results. The first few principal components contain the most information about the data, and they can be used for visualization, dimensionality reduction, or other downstream tasks.

You can also calculate the proportion of total variance explained by each principal component to understand the importance of each in preserving information.

Step 8: Optional - Reconstruction (Inverse Transformation)

If needed, you can reconstruct the data in the original feature space by reversing the projection step. This can be useful for visualizing the data or for further analysis.

In summary, PCA is a technique for reducing the dimensionality of data by finding orthogonal directions (principal components) that capture the most variance in the data. It's a powerful tool for data preprocessing and visualization, often used in exploratory data analysis and feature engineering before applying machine learning algorithms.

Applications of PCA Analysis:

- PCA in machine learning is used to visualize multidimensional data.
- In healthcare data to explore the factors that are assumed to be very important in increasing the risk of any chronic disease.
- PCA helps to resize an image.
- PCA is used to analyze stock data and forecasting data.
- You can also use Principal Component Analysis to analyze patterns when you are dealing with high-dimensional data sets.

Advantages of Principal Component Analysis:

- Easy to calculate and compute.
- Speeds up machine learning computing processes and algorithms.
- Prevents predictive algorithms from data overfitting issues.
- Increases performance of ML algorithms by eliminating unnecessary correlated variables.
- Principal Component Analysis results in high variance and increases visualization.
- Helps reduce noise that cannot be ignored automatically.

Disadvantages of Principal Component Analysis:

- Sometimes, PCA is difficult to interpret. In rare cases, you may feel difficult to identify the most important features even after computing the principal components.
- You may face some difficulties in calculating the covariances and covariance matrices.

- Sometimes, the computed principal components can be more difficult to read rather than the original set of components.

References:

[1] *A step-by-step explanation of principal component analysis (PCA)*. (s.f.). Built In. https://builtin.com/data-science/step-step-explanation-principal-component-analysis

[2] *What is dimensionality reduction - techniques, methods, components - dataflair*. (s.f.). DataFlair. https://data-flair.training/blogs/dimensionality-reduction-tutorial/

[3] *Step-By-Step guide to principal component analysis with example*. (2022, 20 de julio). Hire the World's Most Deeply Vetted Remote Developers | Turing. https://www.turing.com/kb/guide-to-principal-component-analysis#principal-component-analysis-example

[4] G. T. Reddy et al., "Analysis of Dimensionality Reduction Techniques on Big Data," in IEEE Access, vol. 8, pp. 54776-54788, 2020, doi: 10.1109/ACCESS.2020.2980942.

[5] B. M. . Salih Hasan and A. M. . Abdulazeez, "A Review of Principal Component Analysis Algorithm for Dimensionality Reduction", *jscdm*, vol. 2, no. 1, pp. 20–30, Apr. 2021.