Lancaster
University
Leipzig

Course Code: LZPMPC004L: Maths and Stats[1]

Student Id: 39379884

Name: Jingwen Li

1 What is Principal Component Analysis

1.1 Definition

Principal Component Analysis(PCA) is a statistical technique used for dimensionality reduction, which transforms a high-dimension data set into a low-dimension one while retaining as much of the original information as possible. PCA transforms the data into new directions by identifying the direction in which the difference is greatest(the principal component). These new directions are orthogonal to each other(uncorrelated), which preserves the most important structural information in the data while reducing dimensionality.

2 History

Principal component analysis(PCA) is a statistical method, which is used to simply data sets and clarify data structure by reducing dimensions. The development history of PCA can be traced back to the early 20th century.

2.1 Early stage of development

The concept of PCA was first proposed by Karl Pearson, an English mathematician.In the article titled 'On Lines and Planes of Closest Fit to Systems of Points in Space', he described a method to find the best fitting line of a data set by orthogonal projection.

2.2 The foundation of modern PCA theory

In 1933, Harold Hotelling further extended the concept of PCA. The method of Principal Component Analysis he proposed was mathematically equivalent to Pearson's method, but was more systematic and general. Hotelling's work made principal component analysis widely recognized and applied in the field of statistics.

2.3 Improvement of calculation method

In the mid-20th century, application of PCA was promoted by development of computing technology. With the emergence and development of electronic computers, PCA was widely used in practical problems, especially in the fields of data analysis and pattern recognition.

2.4 The expansion of application field

PCA was widely used in various fields from the late 20th century to the early 21st century, such as image processing, gene expression data analysis, financial data, etc. The improvement of computing power and the development of big data technology made PCA an important tool for processing and analyzing high-dimension data.

2.5 Modern improvements

In recent years, with the development of machine learning and artificial intelligence, PCA methods have been continuously improved. For example, variant methods such as Nonlinear Principal Component Analysis(Kernel PCA) and Sparse Principal Component Analysis(PCA) have been proposed to address the analysis needs of complex data sets.

3 Principal component

3.1 What is principal component

Principal components are new variables constructed by linear combinations of original variables. The first principal component(PC1) is the direction in which the variance is greatest in the direction orthogonal to the first principal component, and so on.

3.2 Process of PCA

3.2.1 Data standardization

The first step in PCA is to standardize the data. The aim of standardization is to eliminate scale differences between features. For a data matrix X, each row represents a sample and each column represents a feature.

Suppose there is a data set with **n** samples and **m** features.The data matrix can be expressed as:

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1\mathrm{m}} \\ x_{21} & x_{22} & \cdots & x_{2\mathrm{m}} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\mathrm{n}1} & x_{\mathrm{n}2} & \cdots & x_{nm} \end{bmatrix}$$

Step1, centralization. For each feature j, calculate the average value $\mu_j = \frac{1}{n}\sum_{i=1}^{n} X_{\mathrm{ij}}$ for all samples. And then subtract this average value from each individual feature’s data point.

Step2, calculate the standard deviation $\sigma_{\mathrm{j}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_{ij} - \mu_j)^2}$ of each feature j, which reflects the degree of dispersion of the feature. X_{ij} is the jth feature value of the ith sample of the data matrix.

Step3, use the following formula to get the standardized data.

$$X_{standardized,ij} = \frac{X_{ij} - \mu_j}{\sigma_j}$$

3.2.2 Calculate covariance matrix

After standardizing, calculating the covariance matrix of the data matrix, which represents the linear relationship between features.

The formula of covariance matrix is $\Sigma = \frac{1}{\mathrm{n}-1} X_{\mathit{S}\mathrm{tandardized}}^{T} X_{\mathit{S}\mathrm{tandardized}}$.

The data matrix $X_{\mathit{S}\mathrm{tandardized}}$ obtained according to the standardized.

$$X_{standardized} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1\mathrm{m}} \\ x_{21} & x_{22} & \cdots & x_{2\mathrm{m}} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\mathrm{n}1} & x_{\mathrm{n}2} & \cdots & x_{nm} \end{bmatrix}, \quad X_{\mathit{S}\mathrm{tandardized}}^{\mathrm{T}} = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{\mathrm{n}2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1m} & x_{2\mathrm{m}} & \cdots & x_{nm} \end{bmatrix}.$$

$C = X_{S\tan dardized}^{T} \times X_{S\tan dardized}$, every element $\mathrm{c}_{ij} = \sum_{k=1}^{n} x_{ki} \times x_{kj}$.

After that, Σ is a $m \times m$ square matrix.

3.2.3 Calculate eigenvalues and eigenvectors

After obtaining the covariance matrix, the next step is to solve for the eigenvalues and eigenvectors of the matrix. These eigenvalues and eigenvectors will help to find the principal component direction of the data.

For the covariance matrix Σ , find a set of scalars λ(eigenvalues), and vectors v(eigenvectors), so that the following equation holds: $\Sigma v = \lambda v$. The specific solution processes are as follows:

Step1, solve characteristic equation.The eigenvalue λ of the covariance matrix Σ can be found by solving the following eigenvalue equation $\det(\Sigma - \lambda I) = 0$ .Where **I** is the identity matrix and **det** represents the determinant.

Step2, obtain eigenvalues. By solving the above characteristic equation, the eigenvalues of Σ are obtained: λ_1, λ_2, ..., λ_{m}.

Step3, solve eigenvector. For each eigenvalue, find the corresponding eigenvector by solving the following equation $(\Sigma - \lambda_i I) v_i = 0$. The result is an eigenvector $\mathbf{v}_{\mathbf{i}}$ for each eigenvalue.

3.2.4 Select principal component

The eigenvectors are sorted according to the magnitude of the eigenvalues, and the eigenvectors corresponding to the largest eigenvalues are selected as the principal component direction.

Suppose the data is reduced from **m** dimension to **k** dimension(where k<m). The eigenvector **$\mathbf{v}_1$, $\mathbf{v}_2$, ... , $\mathbf{v}_k$** corresponding to the eigenvalue λ_1，λ_2，...，λ_k are selected as the principal component direction.

3.2.5 Construct transformation matrix

The selected k eigenvalues are arranged vertically to form a $m \times k$ transformation matrix $W = [\mathrm{v}_1 \ \ v_2 \ \ \cdots \ \ v_k]$,where each column is an **m** dimensional eigenvector.

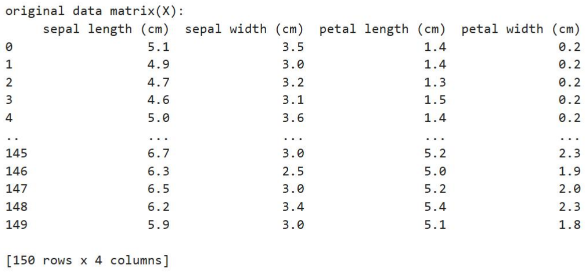
3.2.6 Data transformation and dimensionality reduction

The original data $X_{S\tan dardized}$ into the new k-dimensional space using the constructed transformation matrix **W**. The reduced data matrix $X' = W^T X_{S\tan dardized}$ is obtained, where X' is a $k \times n$ matrix, which means the coordinates of each sample in the new space.

4 Apply PCA to a specific problem

Iris data set is one of the classic data sets in the field of machine learning and data analysis. Use iris data set by processing PCA to distinguish different kinds of iris flowers. This iris data set contains 150 data records, each of which describes four features: sepal length, sepal width, petal length and petal width, and the length of the feature is measured in centimeters. Each record also corresponds to a category label, indicating the iris species to which the record belongs. There are 3 species: iris-setosa, iris-vesicolor, iris-virginica. There are 50 records per category, and the data set is evenly distributed.

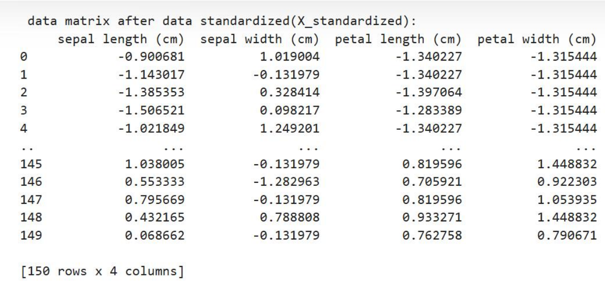
The data is organized into a 150*4 matrix $X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1501} & x_{1502} & x_{1503} & x_{1504} \end{bmatrix}$.



```
original data matrix(X):
     sepal length (cm)  sepal width (cm)  petal length (cm)  petal width (cm)
0                  5.1               3.5                1.4               0.2
1                  4.9               3.0                1.4               0.2
2                  4.7               3.2                1.3               0.2
3                  4.6               3.1                1.5               0.2
4                  5.0               3.6                1.4               0.2
..                 ...               ...                ...               ...
145                6.7               3.0                5.2               2.3
146                6.3               2.5                5.0               1.9
147                6.5               3.0                5.2               2.0
148                6.2               3.4                5.4               2.3
149                5.9               3.0                5.1               1.8

[150 rows x 4 columns]
```

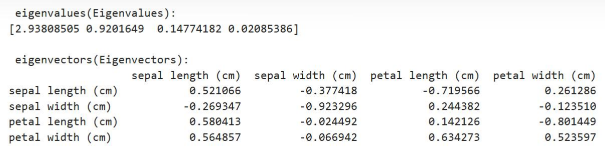
pic1-original data matrix



```
 data matrix after data standardized(X_standardized):
     sepal length (cm)  sepal width (cm)  petal length (cm)  petal width (cm)
0            -0.900681          1.019004          -1.340227         -1.315444
1            -1.143017         -0.131979          -1.340227         -1.315444
2            -1.385353          0.328414          -1.397064         -1.315444
3            -1.506521          0.098217          -1.283389         -1.315444
4            -1.021849          1.249201          -1.340227         -1.315444
..                 ...               ...                ...               ...
145           1.038005         -0.131979           0.819596          1.448832
146           0.553333         -1.282963           0.705921          0.922303
147           0.795669         -0.131979           0.819596          1.053935
148           0.432165          0.788808           0.933271          1.448832
149           0.068662         -0.131979           0.762758          0.790671

[150 rows x 4 columns]
```

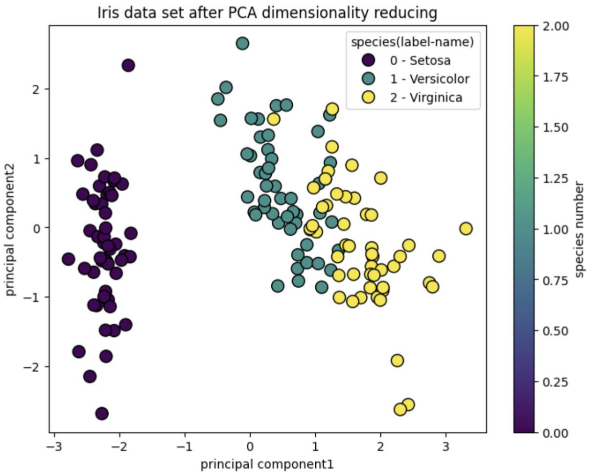
pic2-standardized matrix



```
 eigenvalues(Eigenvalues):
[2.93808505 0.9201649  0.14774182 0.02085386]

 eigenvectors(Eigenvectors):
                   sepal length (cm)  sepal width (cm)  petal length (cm)  petal width (cm)
sepal length (cm)           0.521066         -0.377418          -0.719566          0.261286
sepal width (cm)           -0.269347         -0.923296           0.244382         -0.123510
petal length (cm)           0.580413         -0.024492           0.142126         -0.801449
petal width (cm)            0.564857         -0.066942           0.634273          0.523597
```

pic3-eigenvalues and eigenvectors



Pic4-PCA result

5 Implications of PCA in data science

5.1 Dimensionality Reduction

PCA reduce the number of variables in data sets while retaining most of the original variability. It is crucial when dealing with high-dimensional data, as it helps simplify models, reduce computation time and avoid over-fitting.

5.2 Noise reduction

PCA can help eliminate noise from data by focusing on the components that capture the most variance. It results in cleaner data, which can improve the performance of machine learning models.

5.3 Data visualization

PCA enables the visualization of high-dimensional data in 2D or 3D plots, making it easier to detect patterns, clusters and anomalies.

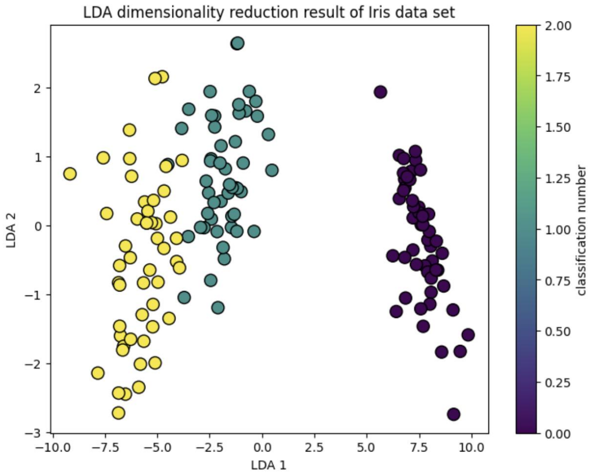
5.4 Feature extraction

PCA can be used to create new features that summarize the information in the original features. These new features can often be more informative for modeling purposes.

6 Alternative techniques

6.1 LDA(Linear Discriminant Analysis)

Comparison with PCA, LDA is supervised and uses category labels to find directions that best differentiate between different categories. This means that LDA generally performs better than PCA in classification tasks.



Pic5-result of LDA