UNIVERSITY OF SCIENCE

FACULTY OF INFORMATION TECHNOLOGY

MULTIVARIATE STATISTICAL ANALYSIS

Report Practice 03
Scikit-learn with PCA and LDA

Lecturers	Lý Quốc Ngọc
	Nguyễn Mạnh Hùng
	Phạm Thanh Tùng
Student	Võ Nguyễn Hoàng Kim
	21127090

CONTENTS

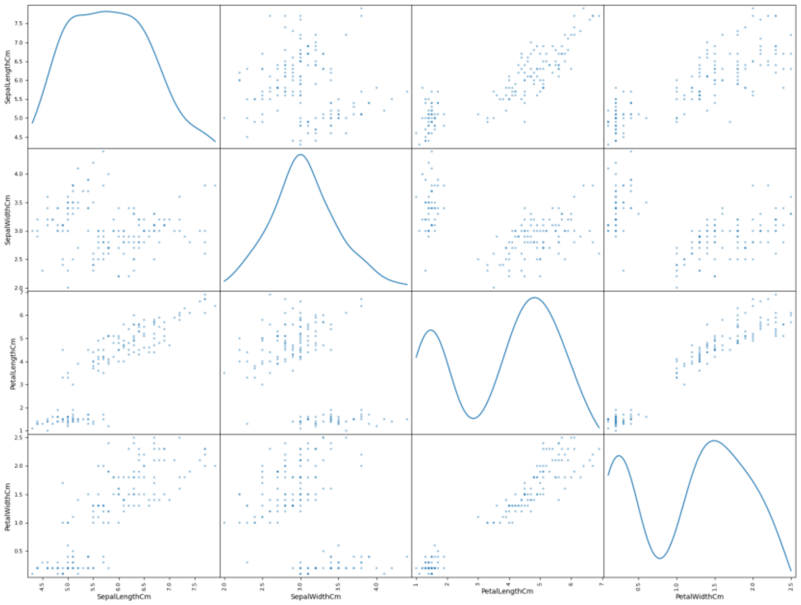
I. SELF-ASSESSMENT FORM

Features		Level of completion
Data	Find a data Iris csv file. Read and describe the dataset.	100%
		100%
Some basic multivariate analysis	With basic statistical quantities. With basic statistical quantities by group. With correlation.	100%
		100%
		100%
PCA and LDA	PCA	100%
	LDA	100%

II. FEATURES

1. Data

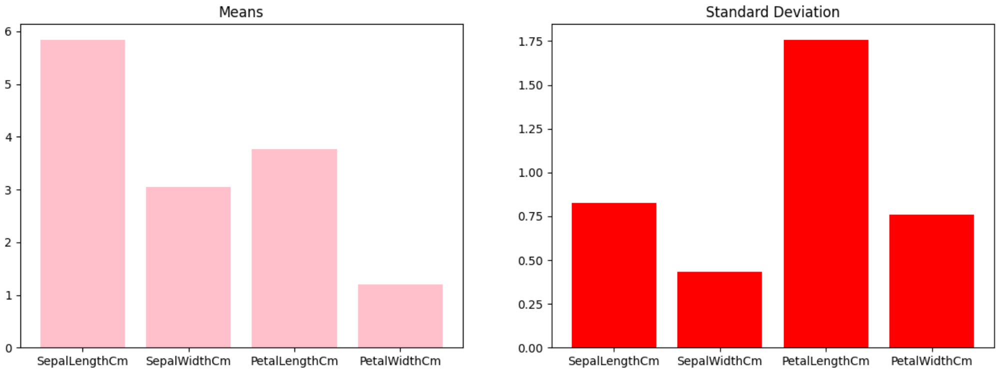
- Download: https://www.kaggle.com/datasets/sachgarg/iris-classification
- The Iris dataset represents 3 kinds of Iris flowers (Setosa, Versicolour and Virginica) with 50 samples each as well as some properties about each flower. One flower species is linearly separable from the other two, but the other two are not linearly separable from each other [1].
- There are 5 attributes in this dataset:
  - SepalLengthCm
  - SepalWidthCm
  - PetalLengthCm
  - PetalWidthCm
  - Species
- The first four columns represent attributes for the size of sepals and the size of petals. The last column is the class identifier (Iris-setosa, Iris-versicolor, Iris-virginica).
- The chart below will provide a comprehensive view of the pairwise relationships and distributions of variables in the dataset:
  - Each off-diagonal cell in the plot matrix represents a scatter plot between two variables from your dataset.
  - These scatter plots show how two variables are related to each other.
  - The diagonal cells of the plot matrix contain kernel density estimates (KDE) for each variable in your dataset.
  - KDE provides a smoothed estimate of the probability density function (PDF) of each variable.



2. Some basic multivariate analysis

a. Basic statistical quantities

- Implement
  - In this section, we will compute the mean and standard deviation of each attribute variable in the dataset.
  - These computations are performed using the **apply()** function from the pandas library with parameters corresponding to the values we want to calculate (here, mean and standard deviation, denoted as **np.mean** and **np.std** respectively).
- Result



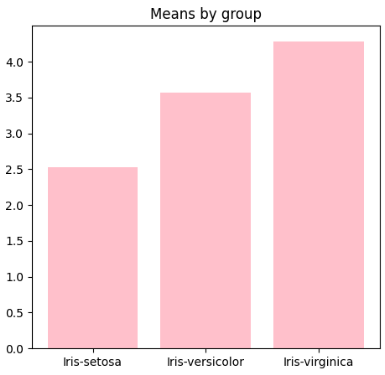
  - The graph illustrates the mean and standard deviation of the attribute elements in the dataset.
  - In the Means graph, it represents the mean values of the attributes across the entire dataset. For each feature element, the graph displays a mean value indicating that the corresponding sample values for that attribute will cluster around the mean value.
    - It can be observed that the mean value of SepalLengthCm is the highest, around 5.8, significantly larger than PetalWidthCm, the feature with the lowest mean value, approximately 1.2. From this, it can be inferred that the values present in the dataset for SepalLengthCm are much larger compared to the other attributes.
  - As for the Standard Deviation graph, it depicts the standard deviation values of each attribute element, which measures the variability or dispersion of the data within that attribute. It

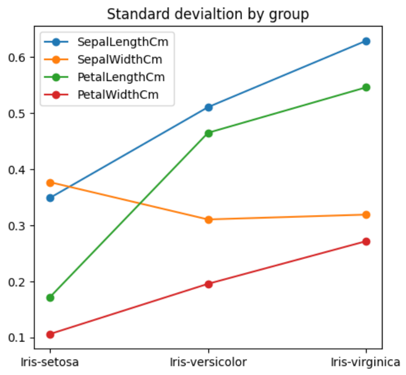
represents the extent of dispersion of values within the attribute element around its corresponding mean value.

- PetalLengthCm is the feature with the largest spread of values around its mean value. This indicates that the corresponding values of this attribute in the dataset exhibit significant variations and unevenness compared to the other features.

b. Basic statistical quantities by group

- Implement
  - In this section, we will compute values such as the mean and standard deviation of variables based on the class labels, meaning calculating these metrics for each Iris flower group (Setosa, Versicolour, and Virginica).
  - First, we group the data by the class identifier. The grouped data is then used to calculate the mean and standard deviation. The resulting output will correspond to each flower group.
- Result

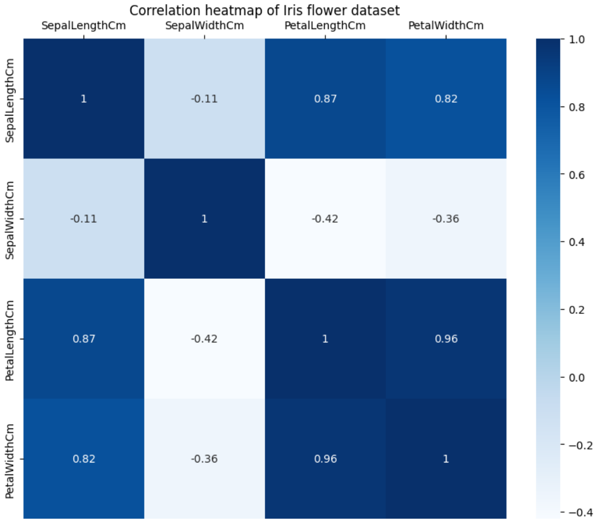




- The graph above illustrates the mean values of the each flower type and the standard deviation of the components within each corresponding flower type.
- In the “Means by group” graph, each column represents the mean value of the data points within a flower type. According to the graph, the mean value of Virginica is the highest, around 4.3, followed by Vesicolor with 3.6, and finally, Setosa has the lowest mean value, approximately 2.5.
- With the "Standard deviation by group" graph, the standard deviation values of the attributes for each flower type are clearly and easily observed. These standard deviation values vary significantly, which means the data grouped by flower types exhibit considerable variability among them. This suggests a pronounced distinction among the different flower types.

c. Correlation

- Implement
  - In this section, we will utilize the **corr()** function, a method available in pandas library, to calculate the correlation matrix. This matrix provides insights into the linear relationship between pairs of variables in the dataset.
- Result



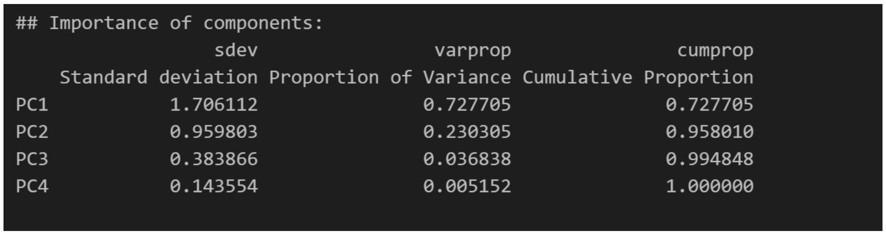
- The heatmap graph represents the intensity of weak to strong correlations between variables in the dataset, where each cell with a intensity of color along with its corresponding value signifies the strength of the correlation between two variables:
  - If a cell is darker in color and its value is close to 1, it indicates a strong positive correlation between the two corresponding variables.
  - If a cell is lighter in color and its value is close to -1, it represents a strong negative correlation between the two corresponding variables.
  - If a cell has no color and its value is close to 0, it means there is no correlation between the two variables.
- With the heatmap, we can identify pairs of variables with strong correlations such as (PetalLengthCm, SepalLengthCm) with 0.87, or (PetalWidthCm, PetalLengthCm) with 0.96. Additionally, we can also determine variables with low correlations like (PetalLengthCm, SepalWidthCm) with -0.42. From these indices, we can identify correlations to evaluate and utilize variables with suitable correlations for computational purposes.

3. PCA and LDA

a. PCA

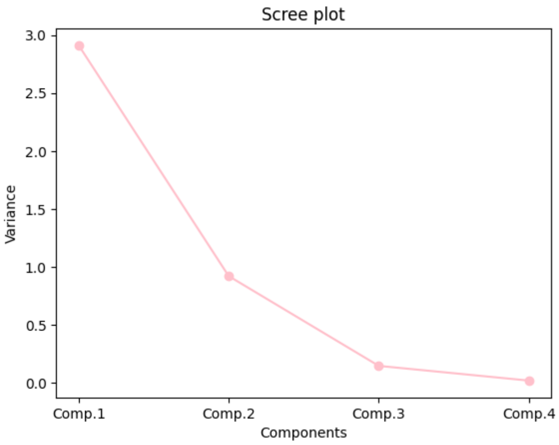
- What is PCA
  - Principal component analysis (PCA) is a linear dimensionality reduction technique with applications in exploratory data analysis, visualization, and data preprocessing [2].
  - The data is linearly transformed onto a new coordinate system such that the directions (principal components) capturing the largest variation in the data can be easily identified [2].
- Implement, Analysis and Result:
  - Before applying PCA to the dataset, it is necessary to preprocess and standardize the data with the aim of removing variables that have too large or too small influences on the analysis or modeling. This is achieved using the **scale()** function.
  - After the data has been standardized, it will be stored in a new variable in the form of a dataframe. PCA will then be applied to the resulting dataframe by **PCA().fit()** function.
  - When we (PCA) to the dataset, each component contains a set of coefficients representing the linear combination of the original features that make up that component. These coefficients indicate the contribution of each original feature to the corresponding principal component.

- After applying PCA to the dataset, we can use the transform() function to project the data onto the principal components, thus enabling the calculation of the mean values of the principal components. Additionally, when applying PCA to the data, we obtain metrics such as Proportion of Variance, which indicates the percentage of variance explained by each principal component, or Cumulative Proportion, which shows the percentage of variance accumulated by all principal components up to the corresponding component. Below is an illustrated result.



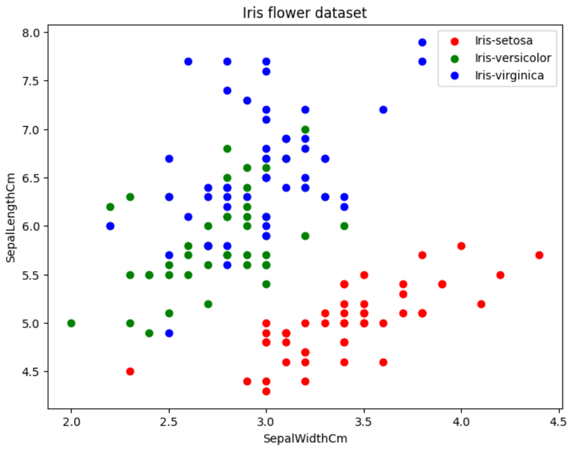
```
## Importance of components:
                  sdev             varprop             cumprop
    Standard deviation Proportion of Variance Cumulative Proportion
PC1           1.706112             0.727705             0.727705
PC2           0.959803             0.230305             0.958010
PC3           0.383866             0.036838             0.994848
PC4           0.143554             0.005152             1.000000
```

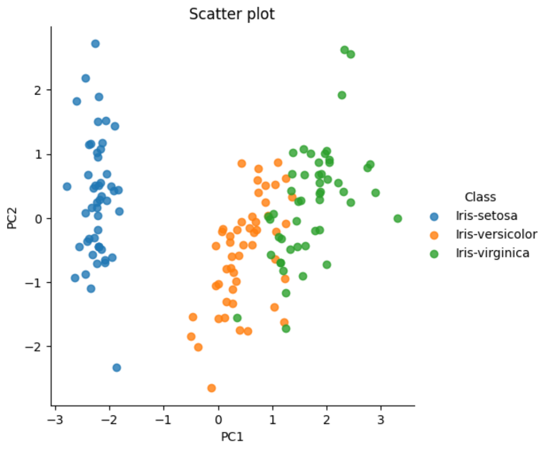
- Additionally, we can calculate the percentage of variance explained by each principal component, aiding in determining the number of principal components to retain while preserving enough variance for the model. Below is a scree plot illustrating this:



    - This plot represents the importance of each principal component in explaining the variability within the data.
    - Each point on the scree plot indicates the variance explained by each principal component. Based on this, we can identify which principal components are important based on the corresponding values on the y-axis for each component. This means that the principal component explains a larger portion of the variability in the data.
    - With this information, we can make decisions regarding the number of principal components to retain in PCA.
- We can plot the data from the two principal components (PC1 and PC2) to visualize the relationship between data samples on these two principal components of PCA (the result is presented below):

Initial Data Plot (1)   After applying PCA (2)

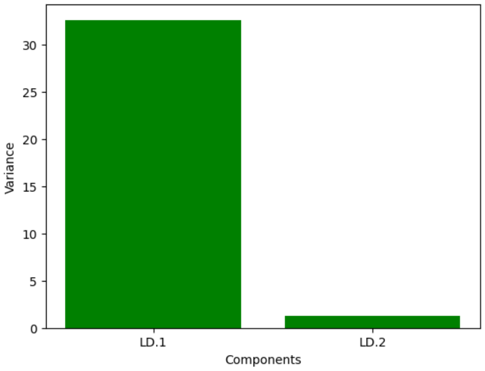




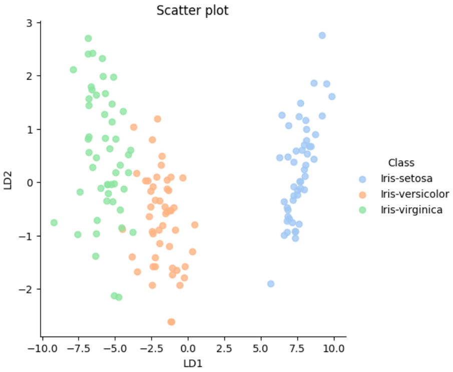
- It can be observed that the data points are distinctly distributed into corresponding clusters for each flower type. Compared to the initially scattered data, the data represented by the two principal components exhibit clear clustering, with data samples close to each other belonging to the same flower type, indicating similarity within each data group.
- Additionally, based on the dispersion of data points in the plot, we can assess the uniformity or differentiation between groups. Specifically, Iris-setosa shows a significant difference in characteristics compared to the other groups. Meanwhile, there is some similarity between Iris-versicolor and Iris-virginica as their data points are relatively close to each other.

b. LDA

- What is LDA
  - LDA is a supervised method of dimensionality reduction that aims to find the linear combination of features that best separates the classes in a dataset. The idea is to reduce the dimensionality of the data while preserving the information that is most relevant for class discrimination [3].
- Implement, Analysis and Result
  - Similar to PCA, before performing LDA on the dataset, we will standardize the data using the **scale()** function. Then, we use the **LinearDiscriminantAnalysis().fit()** function to apply LDA to the dataset.
  - With LDA, its components are typically referred to as "linear discriminants". When we reduce the dimensionality of the data using LDA, we obtain a number of linear discriminants equal to the number of classes in the data minus one (if the number of classes is greater than 2). These linear discriminants are ordered by decreasing eigenvalue corresponding to each linear discriminant. This means that the first linear discriminant retains the largest variance, followed by the second linear discriminant retains the second largest variance, and so on (as the graph below).



- Each linear discriminant contains corresponding coefficients. These coefficients indicate how each input variable contributes to the classification or differentiation between classes in the LDA model. Variables with larger coefficients have a greater influence on classification or differentiation between classes, while variables with smaller coefficients have less influence.
- The following chart represents the dataset of Iris flower after applying LDA to reduce dimensionality.



III. REFFERENCES

[1] SACHGARG, "iris classification," Kaggle, [Online]. Available: https://www.kaggle.com/datasets/sachgarg/iris-classification.

[2] WIkipedia, "Principal component analysis," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Principal_component_analysis.

[3] S. K. Vungarala, "PCA vs LDA — No more confusion!," Medium, [Online]. Available: https://medium.com/@seshu8hachi/pca-vs-lda-no-more-confusion-fc21fb8d06e9.