HO CHI MINH UNIVERSITY OF SCIENCE

MULTIVARIATE STATISTICS ANALYSIS - LAB05

April 20, 2024

Author

Nguyen Viet Kim - 21127333

Contents

1 Introduction

The dataset used in the program is the **Iris dataset** provided by scikit-learn. The dataset consist of f 3 different types of irises' (Setosa, Versicolour, and Virginica) petal and sepal lengths and widths. The purpose of the pratice is to perform CCA to find the canonical variates, and analyze the canonical correlations and loadings to understand the relationships between the two sets of variables.

2 Algorithm Description

The required libraries are Pandas, Matplotlib, Numpy, Seaborn, and Scikit-learn.

```python
# Import required libraries
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
from sklearn.datasets import load_iris
from sklearn.cross_decomposition import CCA
from sklearn.preprocessing import StandardScaler
```

The requirement for the versions of the libraries are specified in the requirements.txt which is included in the source code folder.

2.1 Analyze the correlation between the features of this dataset

The correlation coefficient between the variables are computed as:

$$r_{ik} = \frac{S_{ik}}{\sqrt{S_{ii}}\sqrt{S_{kk}}}$$

The final is the matrix is the skew-symmetric matrix consist of the correlation coefficients of the variables is shown as below:

$$X = \begin{bmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ r_{n1} & r_{n2} & \dots & 1 \end{bmatrix}$$

the matrix can be visualized by the heatmap using Matplolib in the Figure [1].

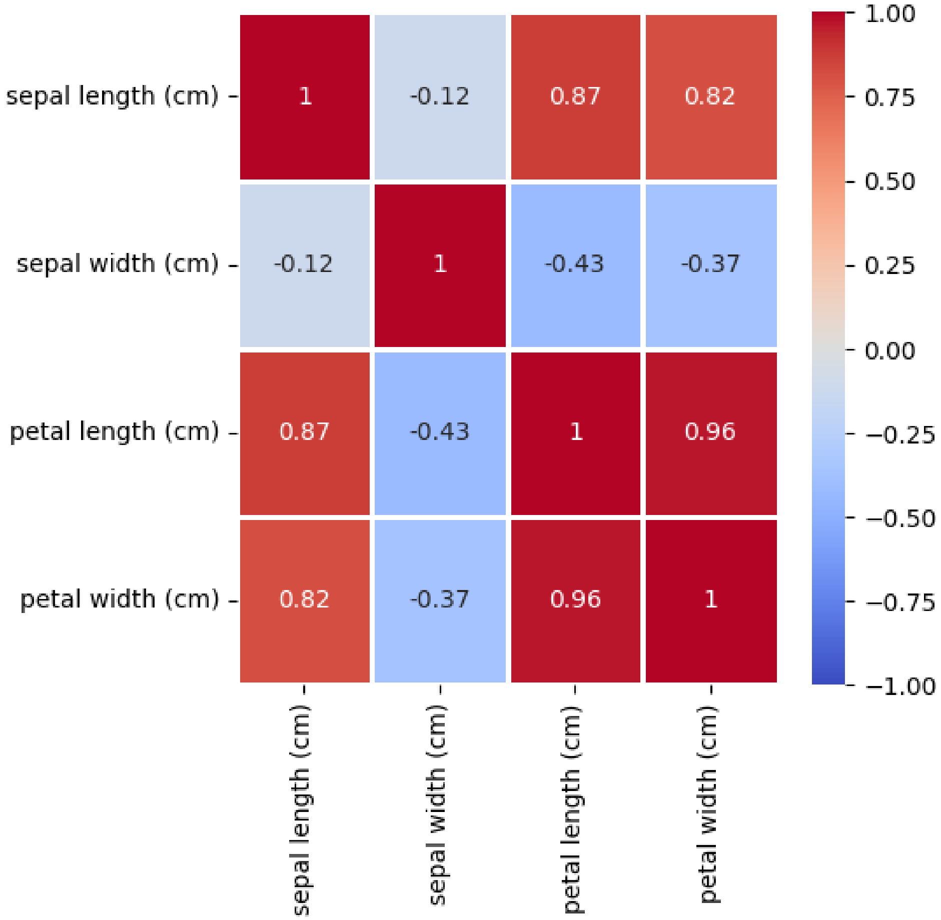

Figure 1: Correlation coefficient heatmap

From the figure, we can observe the correlation between the variables:

- Correlation between sepal width and petal width: 0.75
- Correlation between sepal width and petal length: 0.5
- Correlation between sepal length and petal width: 0.8
- Correlation between sepal length and petal length: 0.7

Sepal width has a moderate to strong positive correlation with both petal width and petal length. This indicates that as sepal width increases, petal width and petal length also tend to increase, but not necessarily in a linear manner. Sepal length also has a moderate to strong positive correlation with petal width and petal length. Similar to sepal width, as sepal length increases, petal width and petal length tend to increase.

2.2 Run CCA using statsmodels, rcca package and skbio

In the context of Canonical Correlation Analysis (CCA), scaling the data ensures that both sets of variables (sepal-related and petal-related) have equal importance in the analysis. This allows CCA to identify the relationships between these sets of variables more accurately, leading to more meaningful canonical variates pairs.

2.3 Check the dependency between canonical variates by correlating canonical variate pairs.

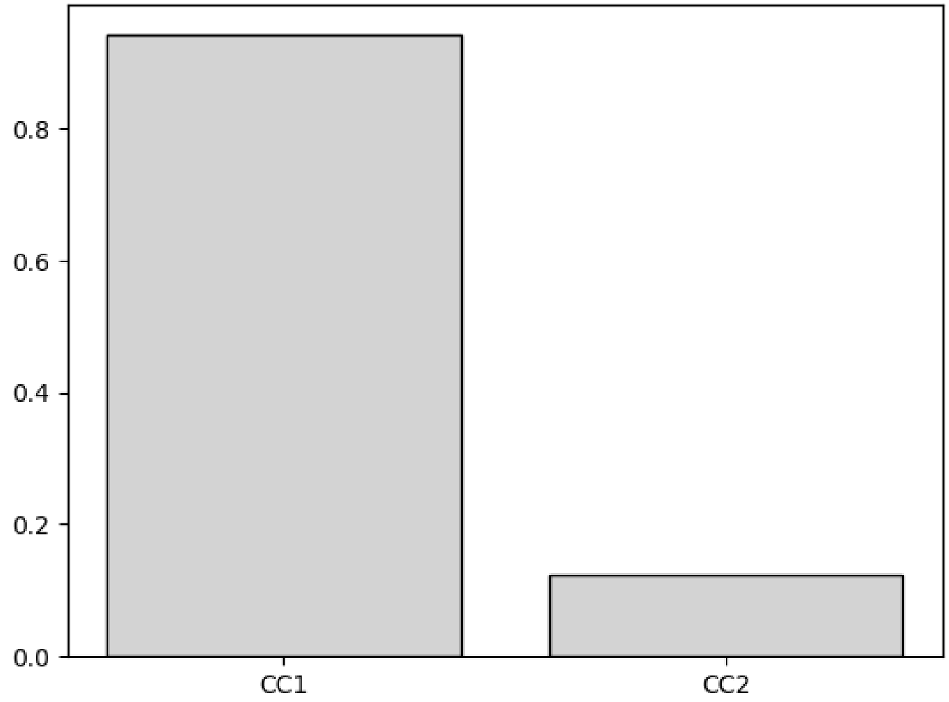

Figure 2: Correlating canonical variate pairs

To determine which pair to be analyzed, we need to take a look at the magnitude of the pairs. These coefficients indicate the strength of the correlation between the canonical variates within each pair. Higher correlation coefficients suggest stronger relationships between the variables represented by the canonical variates.

Therefore, the first canonical variates pair should be analyzed first, as it represents the strongest relationship between the sepal-related and petal-related features.

2.4 Analyzing analyzing the loadings associated with each of the canonical variates

The loadings for canonical varaite of X_1 and X_2 are the following matrices:

$$X_1 = \begin{bmatrix} 0.89224641 & 0.3880084 \\ -0.45786609 & 0.92165584 \end{bmatrix}$$

and

$$X_2 = \begin{bmatrix} 1.5732248 & 0.33270605 \\ 1.45353265 & 0.94303059 \end{bmatrix}$$

For the X_1, The first feature has a strong positive influence on the first canonical variate of the X1 dataset, with a loading of approximately 0.892. The second feature influences the second canonical variate of the X1 dataset, with a loading of approximately 0.921.

For the X_2 dataset, Both features have a strong positive influence on the first canonical variate of the X2 dataset, with loadings of approximately 1.573 and 0.333, respectively. The second feature has a stronger influence on the second canonical variate of the X2 dataset, with a loading of approximately 0.943.

In conclusion, n the X1 dataset, the first feature has a strong influence on the first canonical variate, while the second feature affects the second canonical variate more. In the X2 dataset, both feature have a strong impact on the first canonical variate. However, the second feature seems to have a stronger influence on the second canonical variate.

2.5 Analyzing CCA coefficients

The Figure [3] show the heatmap of the CCA correlation:

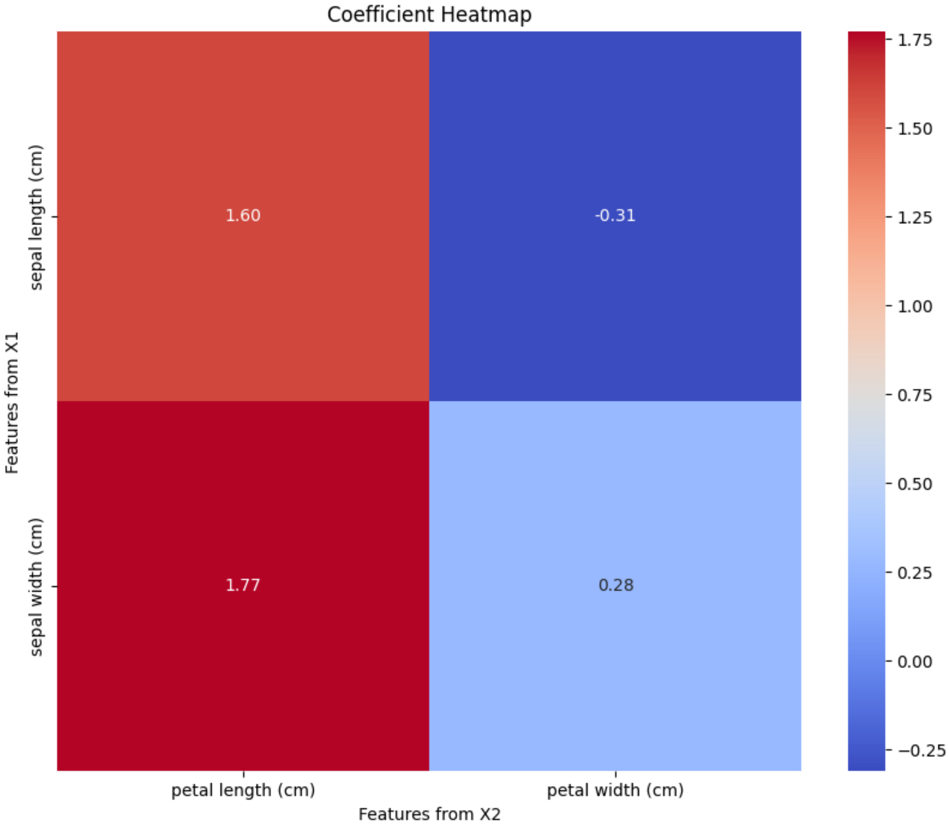

Figure 3: CCA coefficient heatmap

It can be seen that comparing to the heatmap in the second step Fig [1], the number of variables by each axis has decreased to 2 instead of 4. With this map, we can better visualize the correlation between the features of the datasets than in Fig [1]