Homework 6 CSCE 633

Arya Rahmanian

July 2024

Part A

b)

The feature names are sepal length (cm), sepal width (cm), petal length (cm), and petal width (cm)

The target names are setosa, versicolor, and virginica

c)

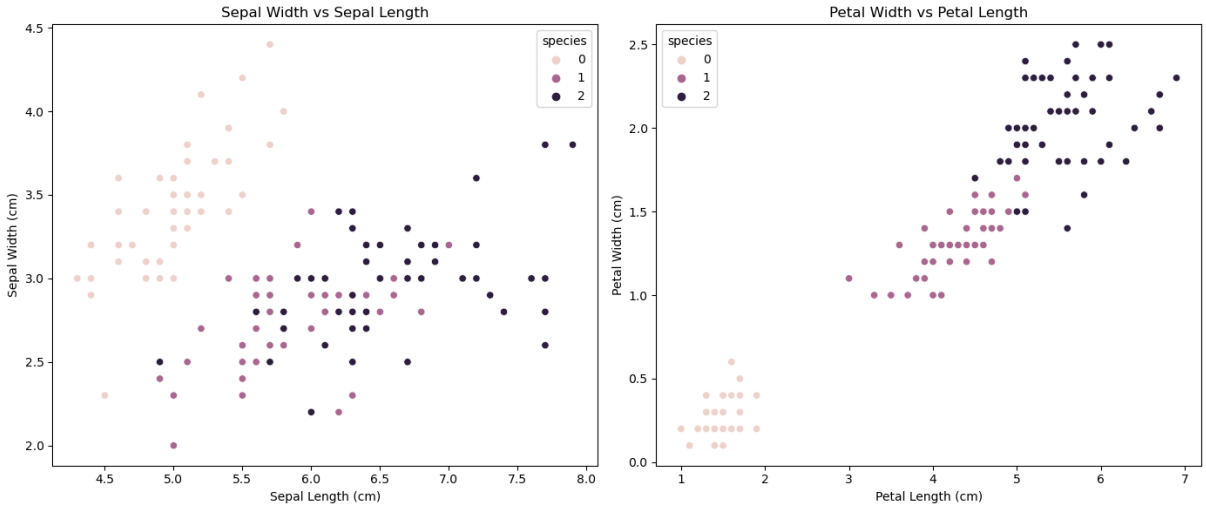

Figure 1: Plots comparing sepal dimensions and Petal dimensions

Are all classes (flower types) easily identifiable based on one of these feature combinations?

Yes. When looking at the petal scatter plot, flowers are pretty distinct and definable. But, when looking at the plot comparing sepal dimensions, setosa is pretty well distinguishable, but versicolor and virginica sepal dimensions are not identifiable.

Are the different flower types distinguished more easily by their sepal or petal features? Which features do you prefer to select?

Petal features are more easily distinguishable, so I prefer to use that to create a model to train on.

Part B: Feature Extraction

2) Check the explained variances and explained variance ratios for each of the three components. What can you infer from these results? Do you still think we need three principal components?

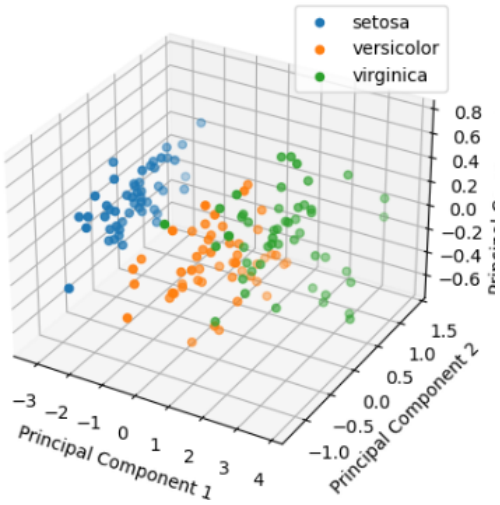

Figure 2: PCA on IRIS Dataset

Output:
Explained variances: [4.22824171 0.24267075 0.0782095]

Explained variance ratios: [0.92461872 0.05306648 0.01710261]

Using three components explain for 99.48% of the total variance. This is great, and one could argue to only use 2 components since that will cover around 97.7% of the variance. But I will continue using 3 components just to ensure maximum coverage. If I wanted to change the number of components, I would continue with 2. Since it covers most of the data with only two components, this may infer that the original features contain a lot of redundancy or correlation. If I needed to do further analysis, only using two components will make visualizing and analysis a lot more simple.

E - Result Analysis) What can you see from the results below? Does PCA actually work? Are the relationships between the chosen words successfully visualized?

Based off the plot, it does look like PCA actually works. The more similar words pertaining to u.s government like u.s, american, military, washington, and officials are closer together. While on the other hand, the nouns pertaining to international economic terms like China, trade, and economic are also grouped close together on their own axis. And lastly, generic government terms such as public, president, country, state, national, and capital and grouped together.The first three components account for about 32.6% of the variance, and the plot reaffirms that value. The plot could be improved using more components, but it would be hard to visualize.

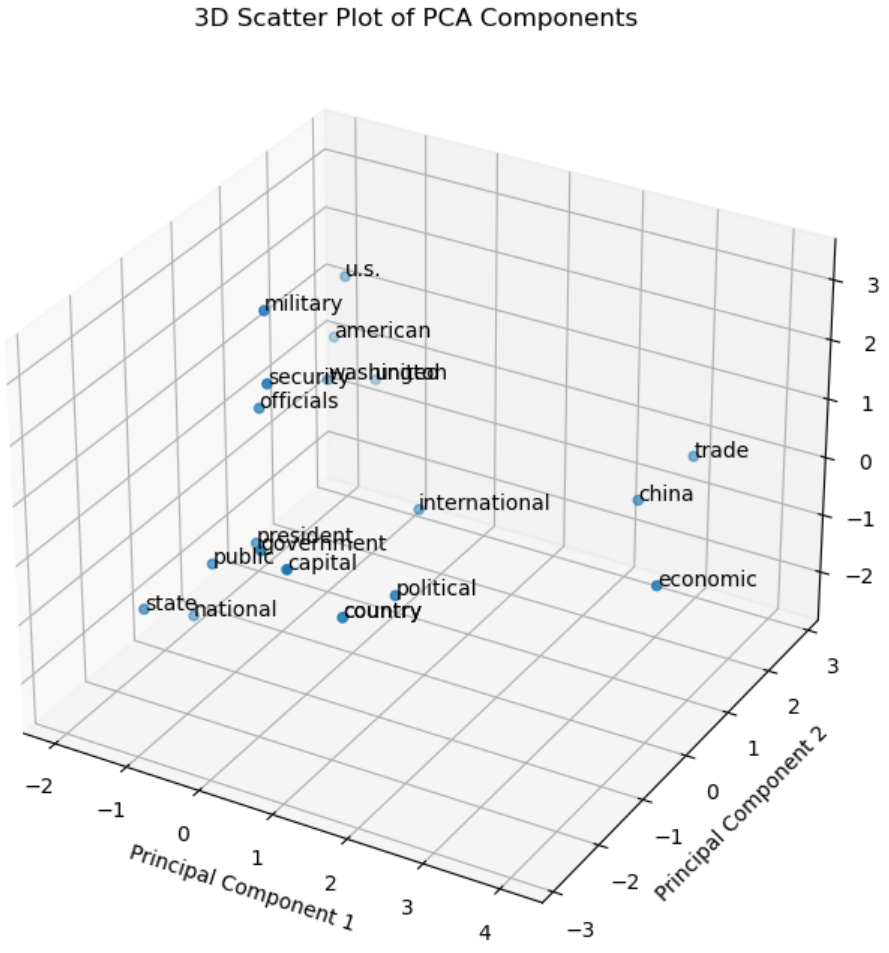

Figure 3: 3D Scatter Plot of PCA Components