Violeta Lio King

Principal Component Analysis and Clustering on Wheat Seed Data

Introduction

The ascension of modern civilization as we know it began when early humans began to develop cultivation practices. Agriculture provided a fundamental shift in how human beings spent their time and thus, provided the means for role specialization and large-scale growth. Perhaps the most important crop that paved the way to this growth is wheat. Over time, many wheat varieties have developed with and without human interaction. In this report, I will examine wheat seed data from three different varieties (Kama, Rosa, Canadian) and investigate their parameters to classify the relationships between different physical characteristics. As surely as the genotypes between different varieties make them distinct from one another, then phenotypes should be able to be used for classification purposes. After all, phenotype is an expression of genotype.

Methods

The original source of the data is from the Institute of Agrophysics of the Polish Academy of Sciences in Lublin but the data was found in UCI's Machine Learning Repository. Using a soft X-ray technique, the internal structure of the wheat kernels were captured and measurements were obtained from these images [1]. There are a total of eight measurement parameters - area, perimeter, compactness, asymmetry coefficient, length of kernel groove, length of kernel and width of kernel.

Due to the number of parameters, the initial step in the analysis was to carry out Principal Component Analysis (PCA) with scaling in order to reduce the dimensionality of the data into something more manageable. The number of principal components (PC) will be decided by considering several factors. The first will be to look at the output and determine how much of the variation is captured by the PCs and if certain features can be omitted to reduce the dimensionality of the data. Once the number of PCs has been chosen, then cluster analysis will be performed. The number of clusters can be determined by several factors, but will mainly rely on the gap statistic to determine the optimal number of clusters. After that, clusters are generated using k-means.

Results

The PCA results indicate that the first three components capture 98.7% of variance while the first two components capture 89% of variance (Table 1).

```
Top features for PC1 (explains 71.87% of variance):
area              0.444474
perimeter         0.441571
widthOfKernel     0.432819
Name: PC1, dtype: float64

Top features for PC2 (explains 17.11% of variance):
asymmetryCoefficient    0.716882
compactness             0.529151
lengthOfKernelGroove    0.377193
Name: PC2, dtype: float64

Top features for PC3 (explains 9.69% of variance):
asymmetryCoefficient    0.679506
compactness             0.629692
widthOfKernel           0.216483
Name: PC3, dtype: float64
```

Table 1. The results of PCA on the seeds data showing the top variables in each PC.

Based on the first three principal components, The first dimension showed a high correlation with area, perimeter, and the width of the kernel (Figure 1). It is not surprising that these specific parameters would have most of the variability. It is evident that area and perimeter are highly correlated to one another and would be captured in the first dimension. The second dimension captures asymmetry coefficient, compactness,and length of kernel groove; the strength of the association is rather weak for the latter two. The third component captures similar variables (asymmetry coefficient, compactness and width of kernel).

All in all, the top features captured by the three principal components encompass all variables in the original dataset except length of kernel. However, looking at overall feature importance based on calculating the absolute weights of the loadings by the explained variance of each component, omitting length of kernel would not be a sound approach as it is an extremely important feature and second only to length of kernel groove (Figure 2). Therefore, the cluster analysis was performed on the original dataset and will be represented by the three principal components.

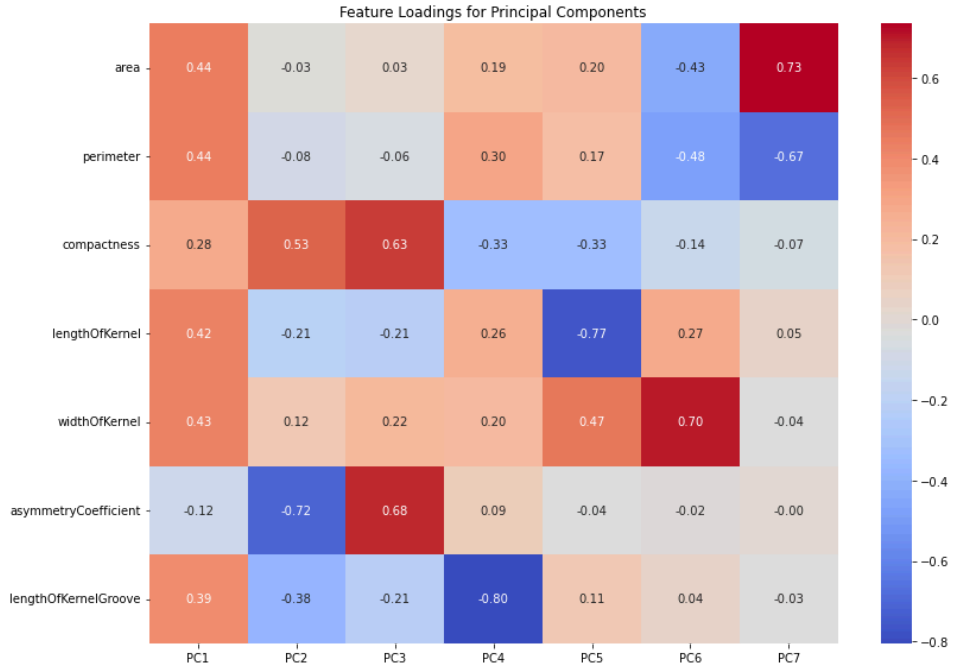

Figure 1. Correlation matrix that depicts which variables contributed to which dimension.

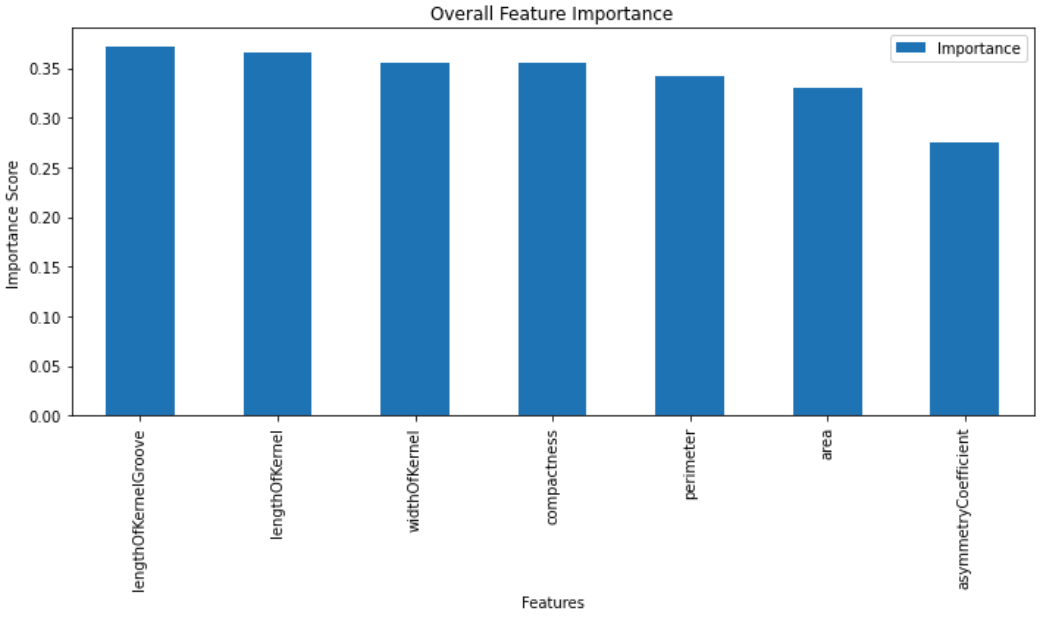

Figure 2. Plot captures overall feature importance for each variable.

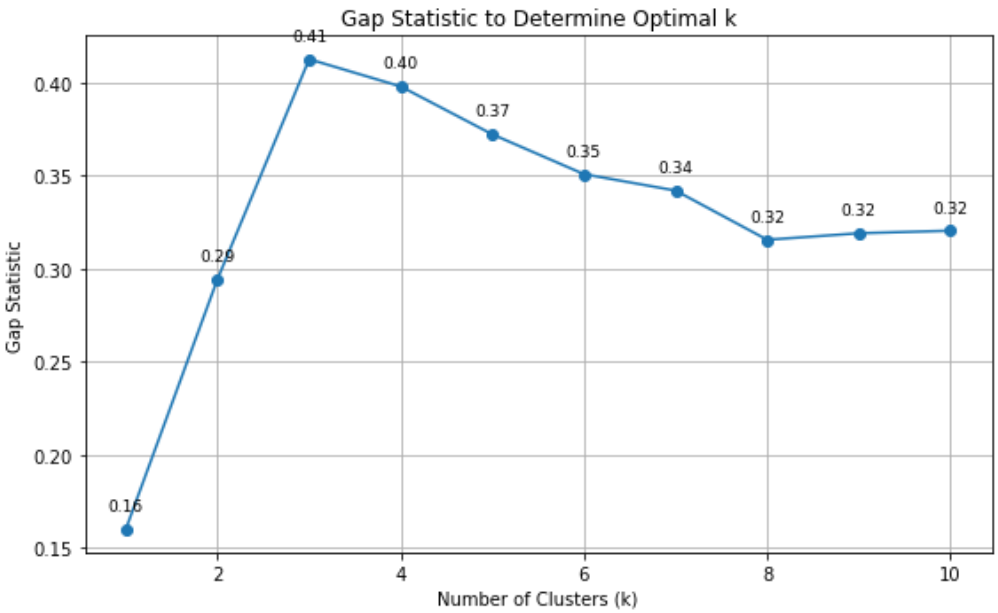

Figure 3. Plot depicts the gap statistic corresponding to the number of clusters.

The gap statistic revealed that the optimal number of clusters is three, which is not surprising considering the fact that there are three rose varieties in the data set (Figure 3). Plotting the first 2 principal components and the three clusters, reveals that there are three distinct clusters but they also overlap slightly (Figure 4). This means that also there are enough unique characteristics to be able to make a reasonable classification of variety, there is still some ambiguity which will lead to misclassification errors. The Canadian variety only overlaps slightly with the Kama variety meaning that it is less likely to have classification errors than the Kama and Rosa varieties. This is even more apparent in the 3d plot (Figure 5).

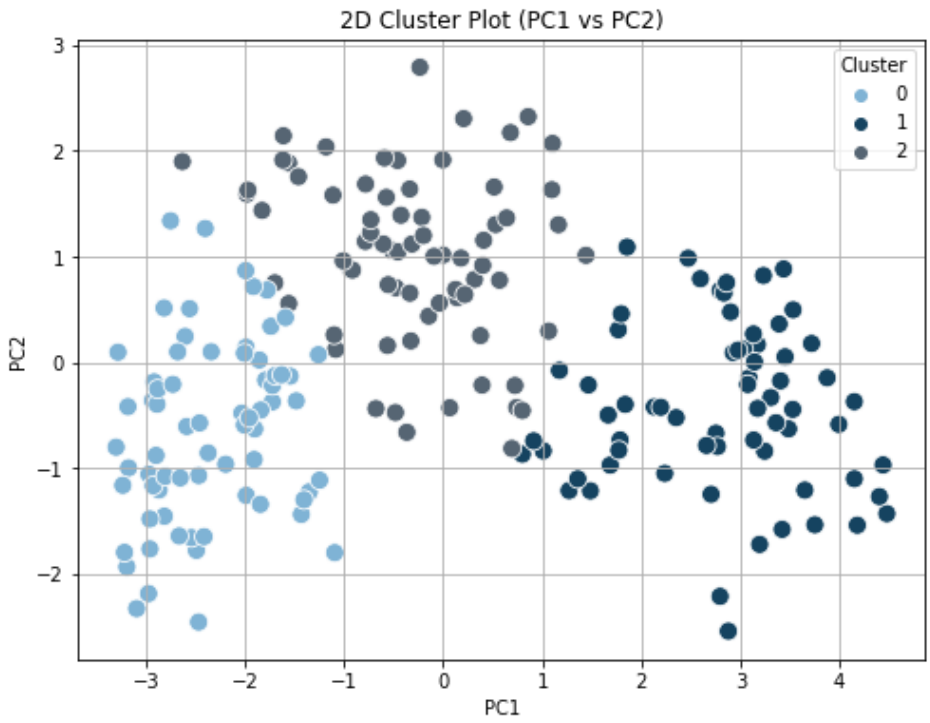

Figure 4. Plot of the three clusters. Depicts a view with only components one and two..

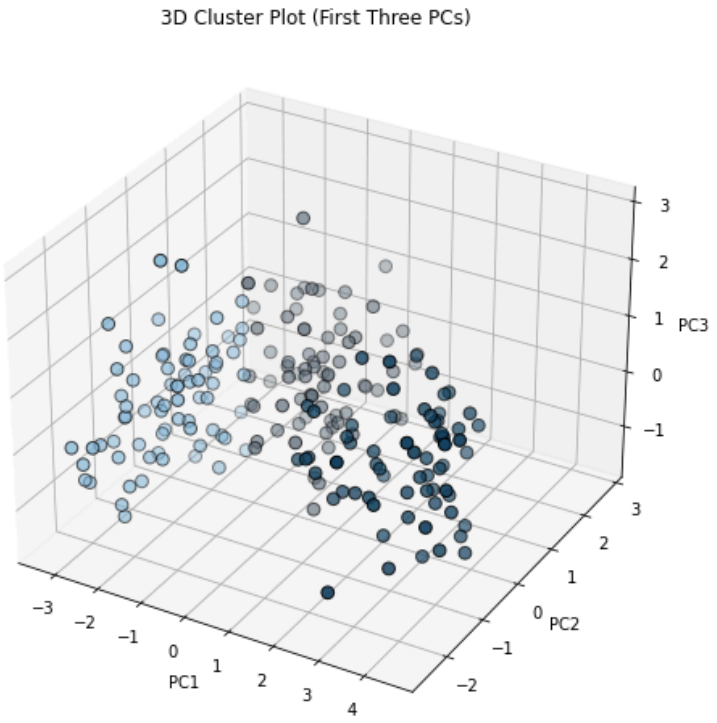

Figure 3. A 3d plot depicting all three clusters with the first three principal components serving as the axes.

Discussion/Conclusion

The cluster analysis plots show that there is some distinction between different varieties of wheat based on all seven parameters. While the clusters themselves appear to be mostly segregated into the correct cluster, there is some overlap between the varieties which indicates that there is some ambiguity. This is further corroborated by the literature which affirm that variation in seed traits exist on a continuous scale rather than a discrete one [6].

While my analysis did not involve genotypes, it is also critical to consider phenotype especially in the context of plant breeding and its role in food security. Since wheat is an essential crop around the world, maximizing its reproductive potential would be highly beneficial especially as climate change becomes an ever present threat to the current viability of crops. Of course, a simple phenotype analysis does not do much good unless it is compared to some metric of fitness, such as identifying which phenotype contributes to the plant's overall health [3, 5]. PCA and clustering has shown that classification can be done and depending on the variety of wheat, can have a reasonably accurate classification rate.

References

[1] M. Charytanowicz, J. Niewczas, P. Kulczycki, P.A. Kowalski, S. Lukasik and S. Zak. A Complete Gradient Clustering Algorithm for Features Analysis of X-ray Images', in: Information Technologies in Biomedicine, Ewa Pietka, Jacek Kawa (eds.), *Springer-Verlag*, Berlin-Heidelberg: 15-24, 2010.

[2] Tchalla Korohou, Cedric Okinda, Haikang Li, Yifei Cao, Innocent Nyalala, Lianfei Huo, Mouloumdèma Potcho, Xiang Li and Qishuo Ding. Wheat Grain Yield Estimation Based on Image Morphological Properties and Wheat Biomass. *Journal of Sensors*, vol. 2020, Article ID 1571936, 2020.

[3] Muhammad Sajjad, Sultan Khan, Muhammad Ashfaq and Wajid Jatoi. Association of seed morphology with seedling vigor in wheat (Triticum aestivum L). *Research in Plant Biology*. 2: 7-12, 2012.

[4] Lisa Sakamoto, Hiromi Kajiya-Kanegae, Koji Noshita, Hideki Takanashi, Masaaki Kobayashi, Toru Kudo, Kentaro Yano, Tsuyoshi Tokunaga, Nobuhiro Tsutsumi and Hiroyoshi Iwata. Comparison of shape quantification methods for genomic prediction, and genome-wide association study of sorghum seed morphology. *PLOS ONE*. 14. e0224695. 10.1371/journal.pone: 0224695, 2019.