MACHINE LEARNING VISUALIZATION BRIEF

Principal Component Analysis (PCA)[1]

Context

The rise of expansive data collection and storage has led to the prevalence of high-dimensional data. Dimensionality, defined as the number of features or input variables within a dataset, may pose a challenge to machine learning when too large. High-dimensional data increases computational costs and may obfuscate learning due to irrelevant inputs or noise. Overfitting arises when an algorithm excessively tunes on the existing data such that it cannot parse between noise and true associations. The resulting model fits training data well but fails to capture real underlying relationships, ineffectively predicting new data.

To reduce overfitting, dimensionality reduction techniques are crucial in the data processing stage. These algorithms transform high-dimensional data to a low-dimensional space to reduce noise while keeping salient information. There are two main approaches to dimensionality reduction: (1) linear methods and (2) non-linear methods or manifold learning. Principal component analysis (PCA) is a linear method that is one of the most widely used dimensionality reduction techniques. It is also useful for unsupervised learning, where instead of predicting output values, models extract patterns from data.

The PCA Method

PCA looks for underlying structures within data by identifying components that explain the most variation in the dataset. PCA transforms a set of correlated variables into a smaller set of aggregated, uncorrelated variables called principal components, while preserving as much of the underlying variation in the original dataset as possible. As such, PCA combines correlated features and removes irrelevant/redundant features.

PCA looks for components that maximize the variance explained within the sample data. Components are linear combinations of the variables computed by adding or subtracting weighted versions of the variables. The algorithm first identifies one component that captures the most variance within the dataset. This component does not represent the entirety of the underlying structure of the data, so the algorithm searches for another component that best captures the remaining unexplained variance, and then iterates the process. In identifying optimal components, PCA attempts to minimize the error that arises from reconstructing the original data using the new features.

A simple demonstration of PCA's dimensionality reduction capability may be gleaned from a sample iris flower dataset which contains information on 4 features (sepal length, sepal width, petal length, and petal width) of 3 types of iris flowers. Figure 1 shows the 4 original dimensions plotted against each other separately. Similar clusters emerge from the pairwise graphs, where Iris-versicolor and Iris-virginica tend to clump together while Iris-setosa tends to be well-separated from the other classes. Figure 2 confirms the observed similarity in clustering formation through the high correlation found among dimensions such as petal length and petal width.

[1] Author: Mary Kryslette C. Bunyi, MS in Data Science for Public Policy Student, Georgetown University

Figure 1: Iris Dataset across Original Features (in cm)

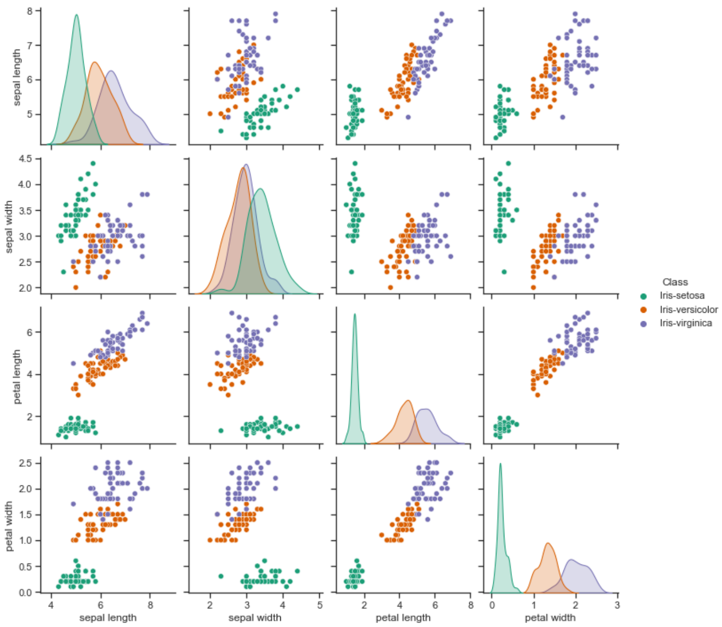
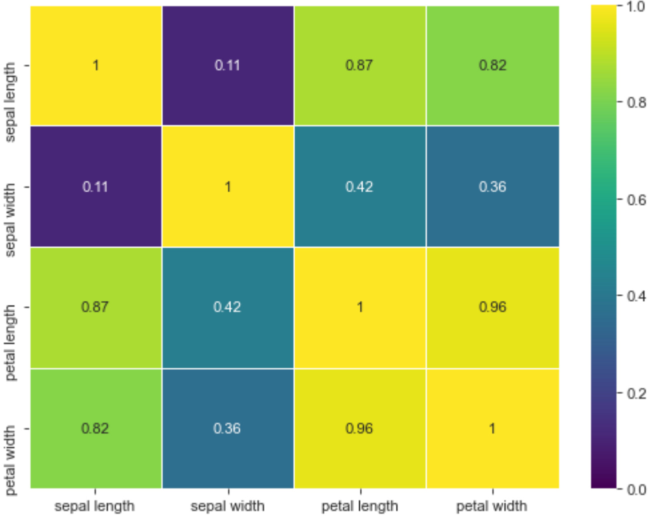


Figure 2: Absolute Correlation between Raw Variables



Applying PCA to the iris dataset reduces the original dataset to a simpler model of 2 features, which together explain 96% of the variance (73% and 23% individually based on Figure 4). Figure 3 shows a plot of the data using the top 2 principal components. The setosa cluster remains distinct from the others, while versicolor and virginica are better separated. Looking at just the x-axis, we could also note that using the first principal component alone can broadly capture inter-variety differences. Setosa occupies the negative x-axis, versicolor tends to concentrate around zero, while virginica tends to take on higher values.

Figure 4 illustrates the decreasing proportion of incremental variance explained as we increase the number of principal components. As mentioned, PCA prioritizes components that account for the largest proportion of variance. This means that PCA implementation requires choosing the number of components that could succinctly but still sufficiently explain variation.

Figure 3: Iris Dataset across Principal Components

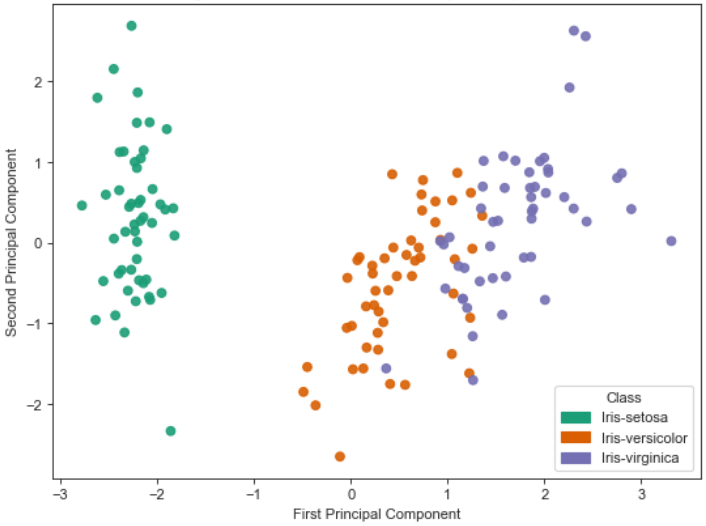


Figure 4: Proportion of Variance Explained by Principal Component

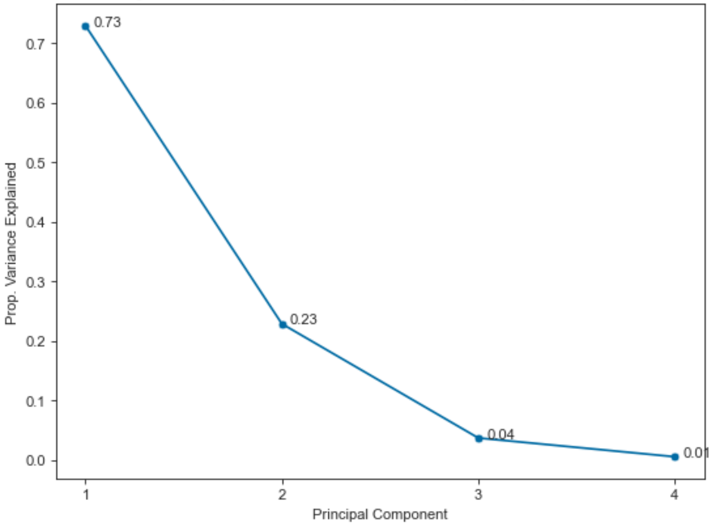


Figure 5 shows the weights (“loadings”) that the principal components assign to the original features. The plot shows that petal width and petal length, which are highly correlated features, are assigned similar weights by the first two principal components. This is an indication that eliminating either feature will likely result in a minimal loss of important information within the data.

Figure 5 likewise shows the importance of specific variables in determining the values of the principal components. The first principal component is highly dependent on sepal length, petal width, and petal length, while the second principal component tends to rely on sepal width and to a lesser extent, sepal length. This result offers an insight into the features that best identify groups of iris flowers and how these features can altogether form a metric for classification.

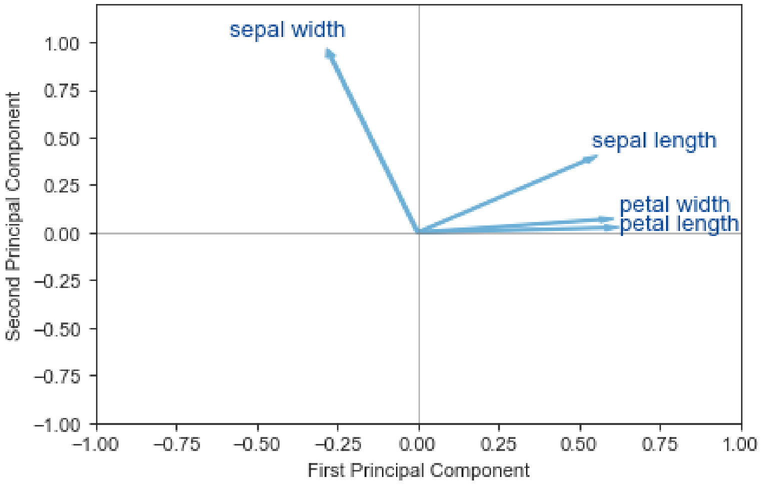


Figure 5: Biplot for First 2 Principal Component Vectors

Conclusion

PCA is a machine learning method which simplifies a set of input variables to reduce computational costs and minimize the risk of systematically incorporating noise into a model. It does so by transforming the original set of features into a smaller set of linear combinations which best capture the underlying variation in the dataset. Use cases for PCA range from exploratory data analysis (investigating patterns within data) to data preprocessing (as a preparation for the implementation of predictive models). Domain knowledge can turn PCA results into conceptual mappings that explain the themes weaving through the components’ most significant features.

Since each principal component is a composite of the original features, the variable transformation that occurs within PCA hampers interpretability. Thus, a crucial step to PCA is determining the number of components that strikes the right balance between capturing as much salient information as possible while minimizing noise.

PCA is essentially the transformation of correlated features to uncorrelated features that most efficiently capture variation within data. Thus, the method is premised on the existence of correlation among variables. Weak correlations will impede the ability of the principal components to capture variability.

PCA’s variance maximization objective in its generation of principal components means that standardization of data is necessary to assign equal importance to all variables. Standardization entails the transformation of each variable’s distribution to achieve a mean of 0 and variance of 1. Without standardization, the principal components will tend to favor variables with larger absolute ranges and values. The focus on variance maximization also makes PCA highly susceptible to outliers. Robust variants of PCA have been developed to address this shortcoming.

References

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Fisher, R.A. "The use of multiple measurements in taxonomic problems" Annual Eugenics, 7, Part II, 179-188 (1936); also in "Contributions to Mathematical Statistics" (John Wiley, NY, 1950).

McCullum, N. (n.d.). Principal Component Analysis in Python - a step-by-step guide. Nick McCullum. Retrieved November 20, 2021, from https://nickmccullum.com/python-machine-learning/principal-component-analysis-python/.

Pramoditha, R. (2021, June 21). Statistical and mathematical concepts behind PCA. Medium. Retrieved November 20, 2021, from https://medium.com/data-science-365/statistical-and-mathematical-concepts-behind-pca-a2cb25940cd4.

Pramoditha, R. (2021, September 28). Principal Component Analysis (PCA) with Scikit-Learn. Medium. Retrieved November 20, 2021, from https://towardsdatascience.com/principal-component-analysis-pca-with-scikit-learn-1e84a0c731b0.