Unsupervised Learning Course Work

Juliana Restrepo Tobar
Universidad EAFIT
Medellin, Colombia
jrestrepot@eafit.edu.co

Olga Lucía Quintero Montoya
Universidad EAFIT
Medellin, Colombia
oquinte1@eafit.edu.co

Abstract—**In this academic report, we present a comprehensive data analysis and description of the Iris dataset. Our approach incorporates a series of unsupervised methods, including pair plots, biplots using Principal Component Analysis (PCA), and Uniform Manifold Approximation and Projection (UMAP). We begin by visually exploring the dataset using pair plots to gain insights into the relationships between different features. Through these analyses, we aim to provide a comprehensive overview of the Iris dataset, shedding light on its inherent patterns and structures. Our findings will contribute to a better understanding of this widely used dataset in order to apply clustering algorithms later on.**

I. Introduction

The history of clustering and unsupervised learning can be traced back to the early days of artificial intelligence research. It emerged as a powerful concept in the mid-20th century, driven by the need to uncover hidden patterns and structures within data without explicit labels or supervision.

One of the foundational techniques in this field is hierarchical clustering, which was introduced in the 1950s. Hierarchical clustering aims to organize data points into nested clusters, revealing a hierarchical structure in the data. However, its applications were limited due to computational constraints and the absence of advanced algorithms.

As computing power advanced, the field of unsupervised learning witnessed significant growth. In the 1960s and 1970s, techniques like K-means clustering and Principal Component Analysis (PCA) gained prominence. These methods allowed for the segmentation of data into distinct clusters based on similarity or dimensionality reduction, respectively.

In more recent decades, with the advent of machine learning and the rise of big data, unsupervised learning has become increasingly important. Clustering algorithms have found applications in various domains, including image segmentation, customer segmentation, anomaly detection, and recommendation systems. Moreover, the field has seen the emergence of sophisticated techniques such as Gaussian Mixture Models (GMMs), DBSCAN, spectral clustering, and UMAP (Uniform Manifold Approximation and Projection), which can handle complex data structures and high-dimensional data.

UMAP, in particular, has gained attention for its ability to perform nonlinear dimensionality reduction and visualization. It has become a valuable tool for exploring and understanding high-dimensional data by projecting it onto a lower-dimensional space while preserving the underlying structure and relationships between data points.

Today, unsupervised learning plays a crucial role in data analysis, feature engineering, and exploratory data analysis. It is a foundation for understanding data distributions, discovering hidden patterns, and enabling data-driven decision-making. Just as in the case of MLPs for neural networks, selecting the appropriate clustering algorithm and parameters is essential for achieving meaningful insights from data. In this project, we delve into the world of unsupervised learning, exploring various clustering algorithms, including UMAP, and their performance on real-world datasets. Our objective is to gain a deeper understanding of the strengths and limitations of these techniques and their impact on data analysis and knowledge discovery.

II. Methodology

A. Data

The Iris dataset is a well-known and frequently used dataset in the field of machine learning and statistics. It was introduced by the British biologist and statistician Ronald A. Fisher in 1936. This dataset consists of 150 samples (n) of iris flowers, representing three different species: setosa, versicolor, and virginica. For each sample, four features (m) are measured: sepal length, sepal width, petal length, and petal width, all in centimeters. The primary purpose of the Iris dataset is to serve as a benchmark for classification algorithms, making it a fundamental resource for tasks like pattern recognition and machine learning model training and evaluation.

B. Descriptive data analysis

1) Basic statistics: In this step, basic statistical measures such as mean, median, standard deviation, and quartiles will be computed for each of the four features (sepal length, sepal width, petal length, and petal width) in the Iris dataset. These statistics will provide a summary of the central tendency and variability of the data.

2) Histograms: Histograms will be generated for each feature to visualize the distribution of data. Each histogram will represent the frequency or count of data points within specified bins or intervals for that particular feature. This will give insights into the data's distribution and any potential skewness. It also serves to assess if there are different populations within the data.

3) Biplot: We will create a biplot by performing Principal Component Analysis (PCA) to reduce the dataset's dimensionality from the original four features (sepal length, sepal width, petal length, and petal width) to just two principal components (PC1 and PC2). This plot allows us to visually assess the contributions of each original feature to the variation in the data and how data points cluster or separate based on these features.

4) Pair plot: For further analysis we'll use a pair plot, sometimes referred to as a scatterplot matrix. This plot will display scatterplots for all possible pairs of features, allowing for the examination of pairwise relationships and correlations within the dataset. This can help identify any strong correlations or patterns between features. This plot differs from the others because we'll use a separate color for each species, which are our target labels.

C. UMAP

UMAP, which stands for Uniform Manifold Approximation and Projection will be applied to the Iris dataset. UMAP is a dimensionality reduction technique used to visualize high-dimensional data in a lower-dimensional space while preserving the essential structure and relationships among data points. The UMAP projection will be created, and the resulting lower-dimensional representation of the Iris dataset will be used for further analysis or visualization.

III. Results

1) Basic statistics: First, we computed a few simple statistics for each feature.

	SepalLength	SepalWidth	PetalLength	PetalWidth
count	150	150	150	150
mean	0.428704	0.439167	0.467571	0.457778
std	0.230018	0.180664	0.299054	0.317984
min	0.000000	0.000000	0.000000	0.000000
25%	0.222222	0.333333	0.101695	0.083333
50%	0.416667	0.416667	0.567797	0.500000
75%	0.583333	0.541667	0.694915	0.708333
max	1.000000	1.000000	1.000000	1.000000

TABLE I: Descriptive Statistics Iris Dataset

2) Histograms: Then, we delved deeper into the distribution of each feature by plotting their histograms.

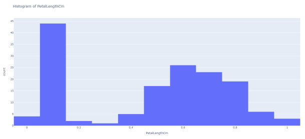

Fig. 1: Petal Length

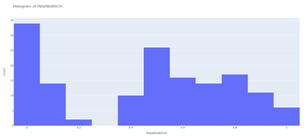

Fig. 2: Petal Width

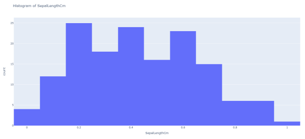

Fig. 3: Sepal Length

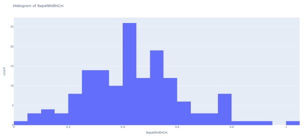

Fig. 4: Sepal Width

3) Biplot: Figure 5 displays a biplot for the data. We can clearly see two clusters.

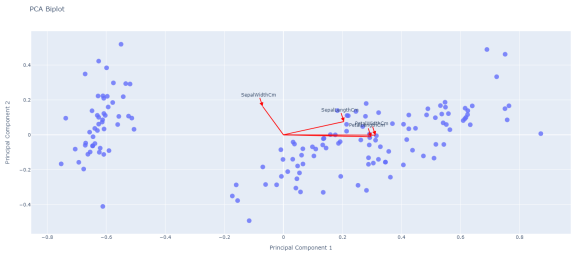

Fig. 5: Biplot

4) Pair plot: Then we took into account the labels and plotted a pair plot or a scatter matrix, shown in Figure 6.

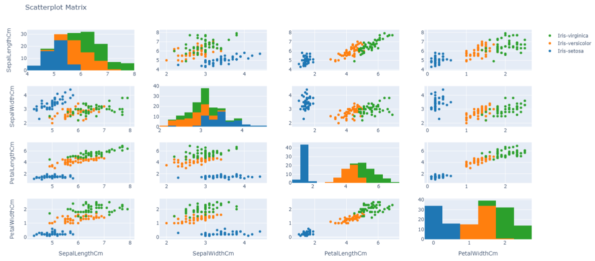

Fig. 6: Pair plot

A. UMAP

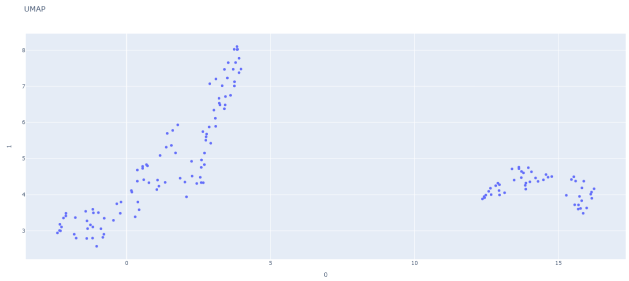

Fig. 7: UMAP

IV. Discussion and Conclusion

The medians and means exhibit a degree of similarity, although certain features display notably high standard deviations. This variance suggests that within these features, the data points are more dispersed. However, the basic statistics alone did not provide sufficient insight to solve the problem at hand. While they offered a summary of central tendencies and variations within the dataset, the presence of high standard deviations and the complexity of the underlying data distribution, as evident in the histograms and other visualizations, indicated that the data's underlying structure and relationships were not adequately captured by these basic statistics.

The histograms representing petal measurements raise suspicions as they resemble normal distributions with low kurtosis. Typically, such distributions hint at the presence of multiple populations within the dataset. A logical approach would be to initially segregate the data by petal lengths and widths before applying clustering algorithms. This approach could prove sufficient with this dataset.

Moreover, the biplot provides clear evidence of at least two distinct clusters within the dataset. Furthermore, it underscores the significance of features related to petals, as indicated by the elongated red lines on the plot.

The pair plot analysis reveals compelling insights about the Iris dataset. Notably, Iris-Setosa exhibits strikingly distinct and well-defined clusters in the joint scatter plots. Even a casual observation of the univariate feature distributions confirms the strong differences between Iris-Setosa and the other two species, Versicolor and Virginica. Conversely, Versicolor and Virginica share some similarities. However, they exhibit a linear separation in nearly all of the joint plots. This linear separation underscores the potential for using these features to differentiate between Versicolor and Virginica effectively. However, it's worth noting that exceptions exist, as the joint plots for sepal width vs. sepal length, and sepal length vs. petal width do not display this same linear separation. These nuances in the pair plot analysis offer valuable insights into the data's structure and provide a foundation for further exploration and analysis in order to distinguish between these two species with precision.

Finally, in the UMAP plot, a striking pattern emerges, confirming the presence of two distinct clusters, mirroring the insights from the earlier biplot analysis. Nevertheless, what sets this UMAP plot apart is its ability to unveil a more intricate structure within the larger cluster. In essence, it helps us see a previously hidden division within the cluster on the left, which gives rise to a total of three discernible clusters, aligning with the classifications proposed by biologists for this dataset.