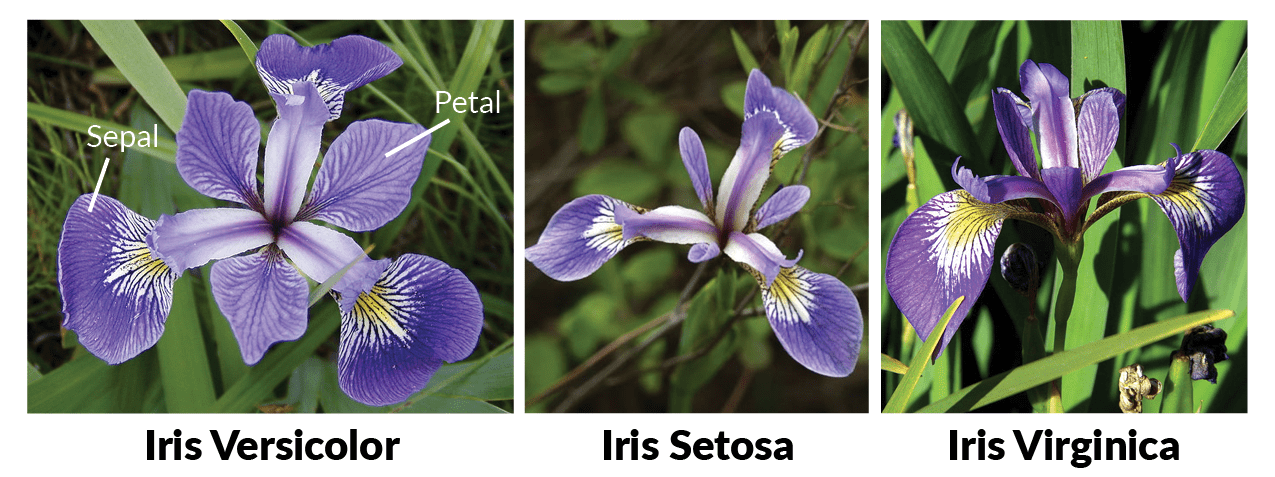
**A Statistical Analysis of Fisher’s Iris Dataset**

by Gerry Callaghan

**Background**

According to wikipedia1, the iris is a flowering plant genus of 310 accepted species with the term iris being widely used as a common name for all Iris species. Irises are perennial plants, growing from creeping rhizomes (rhizomatous irises) or, in drier climates, from bulbs (bulbous irises). They have long, erect flowering stems which may be simple or branched, solid or hollow, and flattened or have a circular cross-section. The inflorescences are in the shape of a fan and contain one or more symmetrical six-lobed flowers (three petals and three sepals).

Source: <https://s3.amazonaws.com/assets.datacamp.com/blog_assets/Machine+Learning+R/iris-machinelearning.png>



**History of Iris Dataset**

The iris dataset dates back almost a century to 1934, having been collected by Edgar Anderson – an American botanist and geneticist. From the Gaspé Pensinsula in Canada, Anderson collected 50 samples from two species of iris, the Setosa and the Versicolor. Anderson measured the width and length of every petal and also the width and length of every sepal. He set out to investigate if the similarities could show one evolved from the other. It should be noted that the current iris dataset contains a third species of iris, the Virginica which differs from the other two samples and was taken from a different colony. It is not clear if indeed Anderson collected this third sample.

Around the time Anderson collected his samples, Sir Ronald Aylmer Fisher, a famous mathematician and statistician, was investigating linear discrimination analysis (LDA)2, that is, a method of finding a linear combination of features that characterises two or more classes of objects or events. The resulting combination could be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification. Anderson’s dataset had characteristics (which we will soon talk about) that made it suitable for Fisher’s study and was therefore used by Fisher in his 1936 classic paper, “The Use of Multiple Measurements in Taxonomic Problems”.

**Appeal of the Iris Dataset**

For Fisher, he needed a reasonable sized dataset, and one could be decomposed into two or more groups, and for there to be a few differentiating features. The subject matter was irrelevant for this study. The iris dataset ticked all the boxes, such as it comprised:

- Three species or classes of iris

- Four differentiating features for each sample were recorded

- Fifty samples of each species were recorded

Therefore, each observation or sample, had five attributes, four of them being measurements (in centimetres) of the width and length of the sepal, and the width and length of petal, while the fifth attribute was its species or class of iris.

More importantly, based on the combination of the four measurement features, one species was said to be linearly separable from the other two, which were not linearly separable from each other.

The idea that one species was linerly separable from the other two made it an ideal sample for Fisher to use in his paper, because he wanted to find the best linear combination of independent variables that will discriminate between the categories of the dependent variables, and determine if significant differences exists among the groups of predictor variables.

**Conclusion**

Fisher showed that the differences between the two species, Setosa and Versicolor was substantially greater than the standard deviations of the compound measurements. Futhermore, the difference between the species, Virginica and Versicolor, was less than four times their standard deviations. Therefore, he concluded that unlike with Setosa, the distributions of the Virginica and the Versicolor are not as easily distinguished from one another based solely on the four measurements.

**Objective of this Analysis**

In my analysis, I wanted to use readily available mathematical and statistical functions in python to import the dataset as a dataframe, manipulate the data into arrays for each of the three class, which I could then use to:

* Calculate the central mean tendencies of each of the four features
* Understand the distributions of those features on histograms
* Plot scatter plots of two features I believed were interdependent
* Determine a regression line for each pair of features
* Deduce the interdependency/correlation of all four features
* Visualize the pairwise relationships using pairplot

I felt that this allowed me to confidently state how each of the of the three species/classes of iris could be distinguised, if at all.

**Conclusion**

According to my statistical analysis, the petal length was the most effective in determining the species of iris, followed closely by the petal width.

I will quickly run through each of the exercises I undertook on this dataset.

1. Summary Statistics  
Sepal Widths -> With a mean width of 3.05 centimetres (cms) for the whole dataset and a standard deviation of 0.43 cms, the mean Setosa sepal width of 3.42 cms was well above the whole sample’s mean while the sepal widths of the Versicolor and Virginica were much lower at 2.77 and 2.97 cms, respectively. The median and the 25%/75% percentiles showed a similar picture.

Sepal Lengths -> The opposite was the case when it came to the sepal length. With a mean length of 5.84 centimetres (cms) for the whole dataset and a standard deviation of 0.83 cms, the mean Setosa sepal length of 5.01 cms and standard deviation of 0.35 cms were well below those of the whole sample. In this case, the sepal lengths of the Virginica at 6.59 cms was much higher than the overall mean, whilethe Versicolor with a mean of 5.94 cms and standard deviation of 0.51 cms was close to the overall mean. The median and the 25%/75% percentiles showed a similar picture.

Petal Widths -> In contrast to the sepal width measurements, the width of the Setosa petal was the smallest mean at 0.24 cms, well below that of the overall mean of 1.2 cms. The Versicolor mean of 1.33 and standard deviation of 0.2 cms was closer to the overall mean while the Virginica at 2.03 cms was well above it. Again, this was mirrored in the median and 25%/75% percentiles.

Petal Lengths -> Similar to the sepal length, the shortest petal was the Setosa with a mean of 1.46 cms compared to an overall of 3.76 cms. Both the Versicolor at 4.26 cms and the Virginica at 5.55 cms were well above the overall mean. This was mirrored in the median and 25%/75% percentiles.

Therefore, already, it could be seen that the Setosa was either well above or below that of the other two, never in the middle when it came to measurements.

2. The histograms of the features were very informative.

Sepal widths -> The histogram of the sepal widths showed a lot of overlap between the three classes, with the Setosa sepal width being higher on average than the other two classes.

Sepal lengths -> This histogram showed a lot of overlap. The Setosa sepal length was at the lower end of the scale, it had very high bars suggesting most of its distribtion was concentrated into only a few bars so its mean would be a lot lower. The Virginica showed the opposite, it would have the higher mean.

Petal widths -> This histogram showed little to no overlap and I deduced from this that the best way to distinguish each class was by the petal width. The Setosa distribution is concentrated in only a handful of bars at the low end of the scale, while the Versicolor and Virginica were at the higher end, but bar 2-3 bars, the Virginica distribution exceeded the Versicolor.

Petal lengths -> This histogram, similar to the petal width, again showed little to no overlap. Futhermore, the classes of iris maintained their positions, with the distribution of Setosa petal lengths concentrated in only a handful of very high bars, signifying very little distribution around the mean. Similar to the petal width, the Virginica had the highest petal lengths with little overlap with the Versicolor, but the main thing is, there was overlap between the two, while the Setosa was much smaller.

3. Scatter plots

Sepal lengths versus sepal widths -> This scatter plot showed a good bit of overlap for the Versicolor and Virginica, but at no stage did any observation of the Setosa overlap with that of either of the other two classes.

Petal lengths versus petal widths -> This scatter plot showed less overlap between the Versicolor and the Virginica, but there was some. In contrast, the Setosa observations were very far away, definitely suggesting it differed greatly from the other two classes.

4. Regression line

Sepal length versus sepal width -> The regression line for the Setosa may have had a relatively similar slope to that of the other two classes, but its intercept was well below that of the other two classes.

Petal lenths versus petal widths -> The regression line for the Setosa differed greatly in terms of slope and especially intercept relative to the other two classes.

5. The correlation

Correlation Coefficient Summary:

* Sepal length and sepal width have little or no correlation, at -0.1.
* Petal length only has a small correlation with sepal width (circa 0.4).
* Sepal length has a relatively large correlation with both petal length and petal width (circa 0.87 and 81, respectively).
* Petal width has a very large correlation with petal length (circa 0.96).

However, given the differences in the summary statistics across the classes of iris, the relationships between the features should be stronger if calculated for each class or iris.

Correlation Coefficient Summary: Setosa  
Relative to the overall iris classification, a massive difference in correlation can be seen between the sepal width and length for Setosa (from -0.10 to 0.75). Likewise there was a large difference for petal length and sepal length (0.87 to 0.26), and petal width and sepal length (0.82 to 0.28). The relationship between the sepal width and petal length goes from a negative 0.42 to a positive 0.17, while that of petal width and sepal width goes from a negative relationship of 0.36 to a positive 0.28. Finally, it can be seen that the relationship between petal length and petal width goes from a massive 0.96 to a much lower 0.31.

Correlation Coefficient Summary: Versicolor  
Relative to the overall iris classification, a massive difference in correlation can be seen between the sepal width and length for Versicolor (from -0.10 to 0.52). Likewise there was a large difference for petal length and sepal length (0.87 to 0.75), and petal width and sepal length (0.82 to 0.55). The relationship between the sepal width and petal length goes from a negative 0.42 to a positive 0.56, while that of petal width and sepal width goes from a negative relationship of 0.36 to a large positive of 0.66. Finally, it can be seen that the relationship between petal length and petal width goes from a massive 0.96 to a more realistic 0.79.

Correlation Coefficient Summary: Virginica  
Relative to the overall iris classification, a massive difference in correlation can be seen between the sepal width and length for Virginica (from -0.10 to 0.45). Likewise there was a small difference for petal length and sepal length (0.87 to 0.86), but large difference in the relationship between petal width and sepal length (0.82 to 0.28). Finally, the relationship between the petal length and petal width is much weaker. The relationship between the sepal width and petal length goes from a negative 0.42 to a positive 0.40, while that of petal width and sepal width goes from a negative relationship of 0.36 to a positive of 0.53. Finally, it can be seen that relationship between petal length and petal width goes from a massive 0.96 to a more modest 0.32.

The large swings in the correlations tells me that the corrleations between the features in the whole dataset are not as reliable as those within the classes.

According to <https://www.analyticsvidhya.com/blog/2024/02/pair-plots-in-machine-learning/>

“A pair plot, also known as a scatterplot matrix, is a matrix of graphs that enables the visualization of the relationship between each pair of variables in a dataset. It combines both histogram and scatter plots, providing a unique overview of the dataset’s distributions and correlations. The primary purpose of a pair plot is to simplify the initial stages of data analysis by offering a comprehensive snapshot of potential relationships within the data.”

* Visualize distributions: Understand the distribution of single variables.
* Identify relationships: Observe linear or nonlinear relationships between variables.
* Detect anomalies: Spot outliers that may indicate errors or unique insights.

A pair plot consists of:

They enable data scientists to:

* Histograms: Diagonal plots showing the distribution of a single variable.
* Scatter plots: Off-diagonal plots showing the relationship between two variables. These can reveal patterns, trends, and correlations.

The seaborn package, which is a high-level visualization package based on matplotlib and integrated with pandas data structures, allows me to create a pairplot of the iris dataset.

Along the diagonal, lies the histograms showing the distributions of that particular feature, be it petal width, petal lengths, sepal widths, or sepal lengths. It can be seen that with respect to the sepal width, the values are all spread across several bars so the variance is high. The distribution is flatter, showing that the numerous samples of each class differ in size and the Setosa sepal width is greater than that of Versicolor or Virginica. With respect to petal length and width, a Setosa differs greatly in size to either the Versicolor or Virginica, the Setosa is much smaller. The one large bar and one small bar tells me that their size is typically that of the tall bar, whereas the other two have more varied sizes.

With respect to the correlations, in the sepal length/petal width box, a positive relationship can be seen across all three classes of iris, starting with Setosa in the bottom left, and working upwards to the top right with Virginica. It must be said, a very similar relationship is shown between the sepal length and the petal length. In the case of the sepal width and petal width, the relationship is almost negative. High values of sepal widths for setosa give low petal widths. Conversely, high petal widths or lengths see small sepal widths. Finally, looking at petal lengths, they look like that have an almost zero relationship with sepal widths or sepal lengths.

In summary, according to the description of the data on UC Irvine (<https://archive.ics.uci.edu/dataset/53/iris>),

one class is linearly separable from the other 2; the latter are not linearly separable from each other.

For me, even a quick look at the pairplot, would tell me that the blue dots for Setosa are almost always situated away from the other two sets of green and orange dots (Virginica and Versicolor). So, I would agree with the description on the website, that one class (Setosa) can be linearly separable from the other two.

Bibliography

1. <https://en.wikipedia.org/wiki/Iris_(plant)>

2 <https://en.wikipedia.org/wiki/Ronald_Fisher>

References

<https://www.angela1c.com/projects/iris_project/the-iris-dataset/>

<https://www.kaggle.com/datasets/uciml/iris>

https://g.co/gemini/share/14f512935c90