Programming and Scripting 2020 Project

Contents

[Abstract 2](#_Toc38483456)

[Introduction 2](#_Toc38483457)

[Materials and Methods 2](#_Toc38483458)

[Results 3](#_Toc38483459)

[Discussion 6](#_Toc38483460)

[Investigation: 6](#_Toc38483461)

[Dataset: 11](#_Toc38483462)

[Conclusion 15](#_Toc38483464)

[References 16](#_Toc38483465)

# 

# Abstract

Using Python libraries imported into a python script to create a summary of each variable that included the Range, Mean, Median, Standard Deviation and the Quartiles. Each species variables displayed with Histograms, Scatterplots and Pair Plot.

The results of this investigation will show how using python to examine the Fishers Dataset can lead to a better and faster analysis.

# Introduction

This project is an investigation into the Fishers Iris Dataset. While this dataset is a corner stone in the data analysis this project will use it to present the usefulness of python in data analysis. This dataset is widely used to teach students techniques in statistics. The dataset is of good quality and without missing fields.

Objectives of this project will be to:

* Explain the benefits of using Python in the analysis of datasets.
* Present the results of data analysis using the Fishers dataset
* Research Fisher’s Linear Discriminant Analysis(LDA)
* To present the code that will show to the user the benefits of ease of use of python in data analysis.
* And to expand knowledge of data analysis and python.

This was accomplished by import various libraries into python using the Visual studio IDE, the project will be version controlled with Git hub. This project will be presented in such a way as to allow a person or group that are not familiar with python to grasp the ease of python when analysing data.

# Materials and Methods

1. Python:
   1. Pandas: this fast, powerful and flexible open source software used for data analysis and manipulation tool. https://pandas.pydata.org/
   2. Matplotlib: Is a library for creating static, animated and interactive visualization in Python. https://matplotlib.org/
   3. Numpy: This library is a fundamental package for scientific computing. https://numpy.org/
   4. Seaborn: Used in Python for data statistical data visualization library based on matplotlib. https://seaborn.pydata.org/
2. Fishers Iris Dataset:
3. GitHub: https://github.com/donalMaher/52167-Programming-Scripting\_Project
4. Running the code: Navigate to where you have download the python script via your command prompt “cmd” run the command “python analysis\_test.py” prerequisite, must have python installed on your computer.



Figure 1 running the code

# Results

The results of call the python script are shown here:

|  |  |
| --- | --- |
| Summary.txt |  |
| Histogram: petal length | Figure 2 Histogram petal length |
| Histogram:  petal width | Figure 3 Histogram petal length |
| Histogram:  Sepal length | Figure 4 Histogram Sepal Length |
| Histogram:  Sepal width | Figure 5 Histogram Sepal width |
| Scatter plot:  Petal length vs Petal width | Figure 6 Petal length vs Petal width |
| Scatter plot:  sepal length vs sepal width | Figure 7 sepal length vs sepal width |
| Pair Plot | Figure 8 Pair plot |
| Code |  |

# Discussion

## Investigation:

The analysis of the dataset will be a compressive summary of the data contained with.

The summary statistics:

There are four key areas is a summary

1. Centrality: the middle value or average
2. Dispersion: how the spread out the values are from the average
3. Replication: how many values there are in the sample
4. Shape: The data distribution.

### Centrality:

The Average is an important measure is statically analysis. It gives an idea about where the data seems to be clustered around. In this summary the mean and the median was calculated.

1. Mean:

|  |
| --- |
|  |

Figure 10 Summary Average mean

Code:

|  |
| --- |
|  |

It can be seen that with two lines of code the mean of the species variable can be displayed.

1. Median:

The mean is average of the values. The median is the measure of central tendency number in a sorted list. The advantage of the median over the mean the median is not skewed so much by extremely large or small values.

|  |
| --- |
|  |

Figure 11 Summary Average media

Code:

|  |
| --- |
|  |

### Dispersion:

How spread out the values are around the average. If the values are close that to the average, and then the iris dataset grouped by species has low dispersion and the opposite for high.

There are several choices to get the dispersion. This summary has the following:

1. Standard Deviation :

This is used when the data is normally distributed. An, average deviation from the mean.

In python it’s not difficult to calculate this.

Standard Deviation:

|  |
| --- |
|  |

This means that the standard deviation is less that 1 standard deviation from the mean.

Code:

|  |
| --- |
|  |

1. Inter-Quartile range (IQR)

Used when the data that is not normally distributed

1. Inter- Quartile range:

|  |
| --- |
|  |

Code:

|  |
| --- |
|  |

1. Range

The range is the difference between the max and the min values

|  |
| --- |
|  |

Code:

|  |
| --- |
|  |

### Replication:

The simples form of summary statistics but it’s important. It’s simply the number of observations.

### Shape

The shape of the data affects the type of summary statistics that best summarize them. The “shape” refers to how the data values are distributed across the range of values in the sample. Generally you expect there to be a “cluster” of values around the average. It is important to know if the values are more or less symmetrically arranged around the average, or if there are more values to one side than the other.

#### Histogram:

The histogram: represents the frequency distribution of data point. Frequency distribution is useful for showing the occurrence of a variable.

In figure 1, the axis y is count and x axis is. In the figures show length and width show no Determining of the species. Even if the data was cluster in the species

Code:

|  |
| --- |
|  |

Figure 9 Histogram code

The scatter plot:

This 2 dimensional plot shows the sepal length and sepal width it’s clear that the blue points are easily determining one of the species, but the other 2 cannot be identified as individual species as they cannot be separated from each other

Linear discriminant analysis:

The dataset contains 150 entries of recorded values which were used to identify the different species of Iris.

Using scatterplot 1 we see the separation of the species. This two dimensional scatter plot shows the three different types of the species based on the features of the flower which we call independent variables or predictors. These are sepal length, sepal width, petal length, petal width. This means there are 4 dimensions.

|  |  |
| --- | --- |
|  |  |

Figure 10 Scatter plot code

Code:

|  |
| --- |
|  |

Pair plots:

|  |
| --- |
| See Figure 8 |

Code:

|  |
| --- |
|  |

Figure 11 Pair plot code

#### 

#### Finding the species type

##### Fishers Linear Discriminant Analysis (LDA)

Fisher developed this dimensional reduction technique to reduce the dimensionality for high dimensional data. This is used for supervised classifications problems.

[6] Fisher developed a linear discriminate model to distinguish the species from each other.

Maximize the distance between means and minimize the variation.

This dimensionality reduction technique is the pre-processing step in machine learning and pattern classification. This is interested on maximizing the separation between the two groups so we can separate into categories which creates a new axis and maximize the distance between means

Minimize the variations (called scatter) (s2) within each category

Where µ1 is the mean of the first variable and µ2 is the 2nd variable this can be summed into distance between the means d

If we apply this to the three species types then we would find the mean of each type and figure out a central point of all the data points and measure the distance (d) between the mean of each type and the central point.

Now the equation is

Using LDA defines new x and y axis which reduces the number of dimensions allowing for separation of the classification of the data. With the species types begin greater than 2 ,LDA creates a plain and lots that data into 3 categories, in this case it’s the species types

The term cluster analysis and the differences between supervised and unsupervised techniques in data mining.

Supervised Learning and Unsupervised Learning:

|  |
| --- |
|  |

These are machine learning tasks. Machine learning is a subset or application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves.

Supervised learning is the process of learning algorithm from the training and dataset.

The mapping of functions to get an output from an input.

This includes:

1. Classification: The placing similar objects into a category.
2. Regression : Used to predict the quantity of a classification category
3. Linear regression: Estimation tool that examines a set of predicts variables make a predicted outcome and which variables are significant.
4. Support vector machine: algorithm that builds a model that assigns new examples to one category or the other

This is used for export systems in image recognition, speech recognition, forecasting, financial analysis and training neural networks and decisions tress.

Unsupervised learning is modelling the structure or the distribution of the data to learn more about it. Where there are no input only outputs. This includes:

1. Clustering or Cluster analysis: Which is a grouping a set of data objects into similar clusters. This groups observations that are similar into the same kind of subsets There are many approaches to cluster analysis. These distance function, used to find similarity between objects. Clustering results are subjective and they dependent on the choices made by the user. This is used to pre-process the data, during exploratory analysis or to pre-train supervised learning algorithms.
2. Association: This is used to set up associations between data objects of large datasets. This technique used to discover interesting relationships between variables in large databases.
3. K-means: This identifies the number of centre points (centroids). This is the averaging of the data finding the centre points. The k-means method will produce k different clusters and are placed as far away as possible, taking the each point belonging to that dataset and associate with the nearest centre point. The when there so no point pending, then positions of k centre point is recalculated, and this method is repeated until the centre point no longer move.
4. Association: The discovering of interesting relationships between variables in large datasets

##### Classification of the species

|  |
| --- |
|  |

Figure 12 the classification of the species

Looking at the species versicolor. It can be seen when trying to differentiate between versicolor species there are some virginica species there are some outliers. These could be described as errors in classification when both species over-lap in the petal length and petal width. To explain this, this could be a mistake in classification of the species type. Both types of species when the petal length gets above 1.6 cm it become difficult to differentiate between both species types. The classification becomes difficult and errors are expected when only two variables are used to determine the species type.

##### Model:

|  |
| --- |
| The model to separate and classify the species could be as simple as:  If(petal\_lenght is less than or equal to 2 and petal\_width is less than or equal to 0.75):  species = setosa  Else if : (petal\_lenght is than 2 and petal\_width is less than or equal to 1.65)  species= versicolor  Else:  species= virginica |

## Dataset:

### About the data set

The Iris flower data set or Fisher’s Iris data set is a multivariate data set introduced by the British statistician and biologist Ronald Fisher in his 1936 paper. This is a very famous and widely used dataset by everyone trying to learn machine learning and statistics. The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimetres. The fifth column is the species of the flower observed.

|  |
| --- |
|  |

Figure 12 species of flower in dataset

The dataset contains samples of three species of Iris:

1. Iris setosa

2. Iris virginica

3. Iris versicolor

The four features/independent variables that were measured to determine the species are:

1. Sepals length

2. Sepals width

3. Petals length

4. Petals width

There are 150 entries in total in the dataset

# Conclusion

* Python

Using python and it’s in libraries to display data is relativity easy and to help visual datasets.

Using the describe function allows the dataset to be summarized in just a few lines of code.

|  |
| --- |
|  |

Figure 13 describe function

|  |
| --- |
|  |

In python creating a very good summary of the dataset is easy, once you research how to do it.

If you wish to separate the summary into individual components such as mean, median and range. This also can be done with relative ease. Python makes the summarisation of datasets to be simple coding task.

The histograms while a great visual tool created to get the mean value on the variables of the dataset. Visualisation of the mean is a very handy tool when describing the LDA. In this context of classification the variables using the histogram can show the mean value of petal length verses petal width and sepal length verses sepal width. The histograms shown in figures 2,3,4,5 just show the count of the entire petal width or petal length. Determining the different species on this mean alone using a histogram in my option cannot be done.

Scatterplots give a better representation than Histograms in this case. And from the scatterplot figure 6 there is better chance to create a model for the dataset. Using the scatterplot classification of the different species can be seen in figure 12. This is also seen in the pair plot. The par plot is gives move insight in the dataset.

This project gives insights into various machine learning concepts supervised and unsupervised learning, Fishers LDA and PDA. While I had attempted to code LDA into this project I failed to get find a reasonable an uncompleted and simplified example of this technique in python even though I found a very good blog on the topic [7] and settled on describing the process. And will be looking forward to diving further into this topic in the coming semesters.

# References

|  |  |
| --- | --- |
| 1 | Python, p, (2020), statistics — Mathematical statistics functions — Python 3.8.2 documentation, https://docs.python.org/3/library/statistics.html,(23/03/2020) |
| 2 | Iris flower data set, <https://en.wikipedia.org/wiki/Iris_flower_data_set>,( 04/03/2020) |
| 3 | Kiteq,k,(2020), How to label axes in a Seaborn Bar Plot in Python, https://kite.com/python/answers/how-to-label-axes-in-a-seaborn-bar-plot-in-python , (16/04/2020) |
| 4 | Tutorialspoint.com,T(2020),Seaborn-Facet-Tutorialspoint,h ttps://www.tutorialspoint.com/seaborn/seaborn\_facet\_grid.htm,(11/04/2020) |
| 5 | Manisha sirsat,MK,2020, Data Science and Machine Learning : Exploratory Data Analysis (EDA), https://manisha-sirsat.blogspot.com/2019/12/exploratory-data-analysis.html,(17,04,2020) |
| 6 | Linear discriminate analysis, https://en.wikipedia.org/wiki/Linear\_discriminant\_analysis,(19/04/2020) |
| 7 | Python LDA (2014), python LDA, <https://sebastianraschka.com/Articles/2014_python_lda.html>, (17,04,2020) |
| 8 | Jake vanderplas, 2017, *Python data science handbook*,1st,O’Reilly, Sebastopol |
| 9 | https://www.youtube.com/watch?v=cpZExlOKFH4  https://youtu.be/b7JuBsswDlo |