Introduction to the Iris dataset

The Iris dataset was one of the first, widely available machine learning test datasets and features labelled data for 3 species of flowing plant (*Iris Setosa*, *Iris Versicolor*, and *Iris Virginica*) with attributes; sepal length, sepal width, petal length and petal width. Using these attributes alone it is possible to distinguish between the different iris species and train a classifier to detect each. Whilst this may not have the most meaningful commercial implications nor be the most complex example of machine learnings power it is simple and allows for a clear outline of the limitations and strategies employed by many machine learning algorithms.



Fig 1. Iris Setosa example

A limited dataset is available from sklearn containing 150 samples complete with attributes and classes. The dataset contains 50 examples of each flowing plant who have generalised statistics shown below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sepal length (cm) | Sepal width (cm) | Petal length (cm) | Petal width (cm) |
| Iris setosa | 5.0 (4.3, 5.8) | 3.4 (2.3, 4.4) | 1.5 (1.0, 1.9) | 0.2 (0.1, 0.6) |
| Iris versicolor | 5.9 (4.9, 7.0) | 2.8 (2.0, 3.4) | 4.4 (3.0, 5.1) | 1.3 (1.0, 1.8) |
| Iris virginica | 6.5 (4.9, 7.9) | 3.0 (2.2, 3.8) | 5.6 (4.5, 6.9) | 2.0 (1.4, 2.5) |

Table 1. Iris dataset statistics shown median (min, max)

The Iris dataset is so common because it shows clearly the limitations associated with non-linearly separable problems. This is visually shown in Fig 2a-f. below:

![A close up of a map

Description automatically generated]()![A screenshot of a cell phone

Description automatically generated]()![A screenshot of a cell phone

Description automatically generated]()

![A close up of a map

Description automatically generated]()![A screenshot of a cell phone

Description automatically generated]()![A close up of a map

Description automatically generated]()

Fig 2a-f. scatterplots of the attributes of the Iris dataset.

Due to the fact that the dataset is widely available and well known, it creates great foundation for the exploration of methods in framework which is well understood. It is expected that *Iris Setosa* is easily distinguishable from other classes given the liner separability shown in Fig 1a-f. *Iris Versicolor* and *Iris Virginica* overlap on many of the attributes and hence most of the error in classification problems will arise from misclassification of one as the other.

Code:

﻿import numpy as np

from sklearn.datasets import load\_iris

import matplotlib.pyplot as plt

import statistics

data, classifier = load\_iris(return\_X\_y=True)

statistics.median(data[np.where(classifier==2)][:,3])

min(data[np.where(classifier==2)][:,3])

max(data[np.where(classifier==2)][:,3])

species = ['Setosa','Versicolor','Virginica']

for i in range(0,3):

plt.scatter(data[np.where(classifier==i),2],data[np.where(classifier==i),3], label=species[i])

plt.legend()

plt.xlabel('Petal length [cm]')

plt.ylabel('Petal width [cm]')

plt.title('Petal length v Petal width within the Iris dataset')

plt.show()

References:

[1] https://www.plant-world-seeds.com/images/item\_images/000/007/023/large\_square/iris\_baby\_blue.jpg?1500653527