Iris Dataset Analysis

Francis Alvarez

5/14/2018

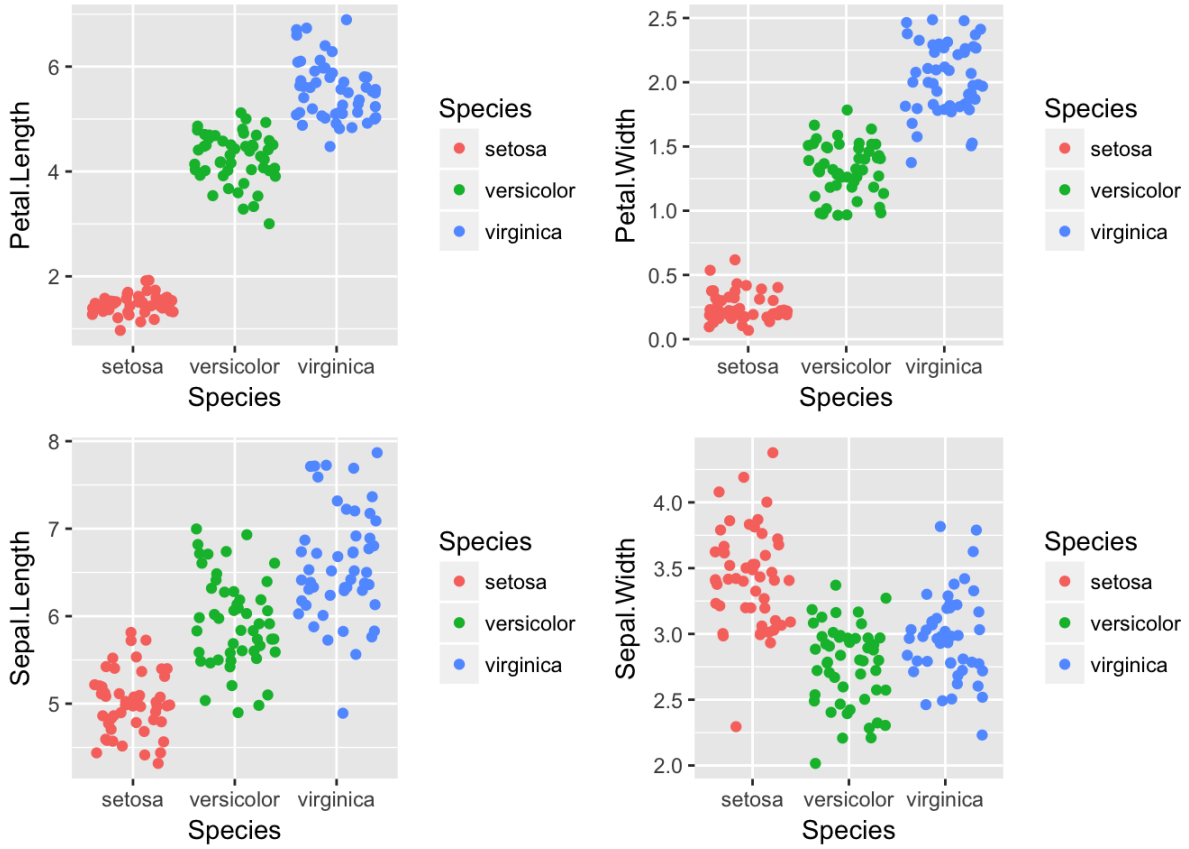
Purpose:

The goal is to gain experience analyzing data sets with varying statistical and machine learning techniques. We are working with the Iris dataset. There are four different traits that are recorded for three different types of species. Each species has 50 samples for a combined total of 150 samples. For this specific data set we will focus on differences among the different species and later see if we can correctly classify each species based on the numerical attributes of the sample.

We begin by plotting the data set to get a quick representation of the values and distribution among species. The structure of the dataset, as well as a sample header, can be seen below.

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```



Mean Across Species

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
setosa	5.006	3.428	1.462	0.246
versicolor	5.936	2.770	4.260	1.326
virginica	6.588	2.974	5.552	2.026

Standerd Deviation Across Species

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
setosa	0.3524897	0.3790644	0.1736640	0.1053856
versicolor	0.5161711	0.3137983	0.4699110	0.1977527

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
virginica	0.6358796	0.3224966	0.5518947	0.2746501

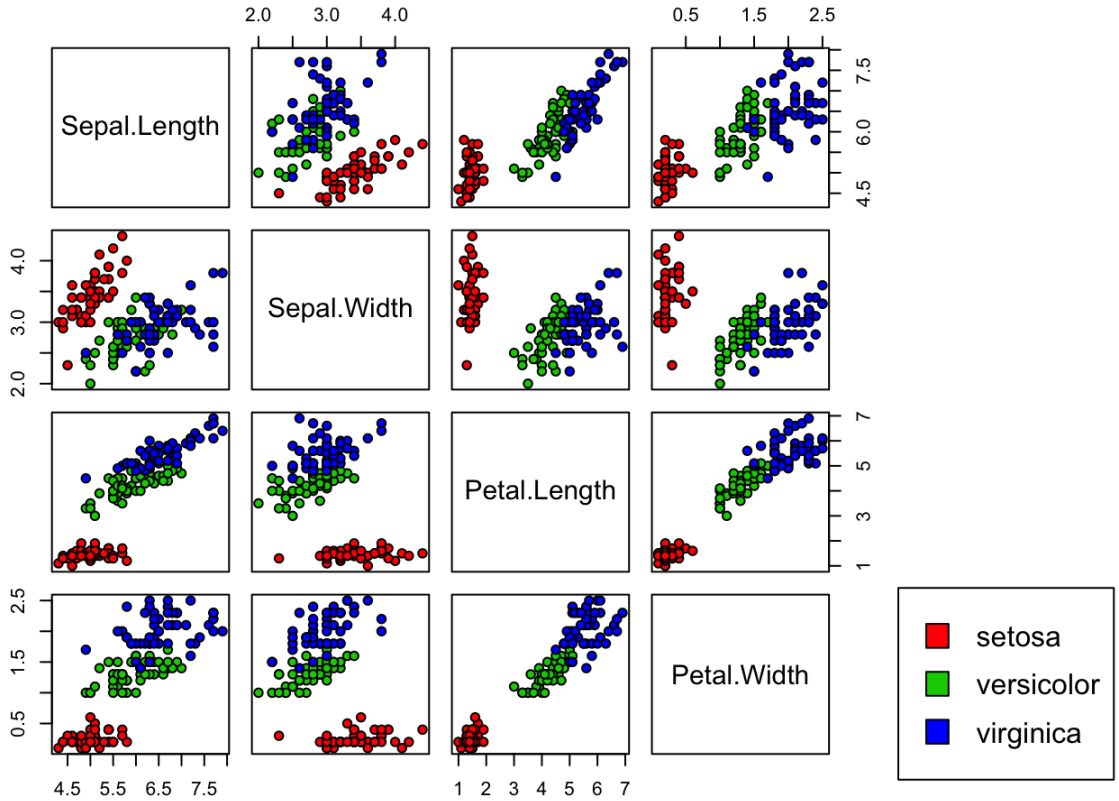
We want to check whether the means across each species are equal. We initially focus on petal length.
ANOVA F Test
$H_0 : \mu_1 = \mu_2 = \mu_3$ versus H_1 : not all the μ_j's are equal

```
oneway.test(Petal.Length~Species)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  Petal.Length and Species
## F = 1828.1, num df = 2.000, denom df = 78.073, p-value < 2.2e-16
```

There is significant evidence to reject that the mean petal length across each species are equal. We could continue across other traits of flowers but we now will focus on classification of species.

Correlation Among Each Treatment



From the plots above we can see the correlation among features. As a whole, petal length and petal width appear to be highly correlated, but once we look at the values as a species there is not much correlation. In certain instances, some species have high correlation for a specific two features while another species will not, for example, petal length versus sepal length.

The values below are ordered by: setosa, versicolor and virginica.

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000   0.7425467    0.2671758   0.2780984
## Sepal.Width     0.7425467   1.0000000    0.1777000   0.2327520
## Petal.Length    0.2671758   0.1777000    1.0000000   0.3316300
## Petal.Width     0.2780984   0.2327520    0.3316300   1.0000000
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000   0.5259107    0.7540490   0.5464611
## Sepal.Width     0.5259107   1.0000000    0.5605221   0.6639987
## Petal.Length    0.7540490   0.5605221    1.0000000   0.7866681
## Petal.Width     0.5464611   0.6639987    0.7866681   1.0000000
```

```
##              Sepal.Length Sepal.Width Petal.Length Petal.Width
## Sepal.Length    1.0000000   0.4572278    0.8642247   0.2811077
## Sepal.Width     0.4572278   1.0000000    0.4010446   0.5377280
## Petal.Length    0.8642247   0.4010446    1.0000000   0.3221082
## Petal.Width     0.2811077   0.5377280    0.3221082   1.0000000
```

Species Classification

The next focus is being able to correctly classify each species simply by the listed four traits. We will utilize three different methods: z-score classification, k-means clustering and neural networks.

Classification by Z-Scores

This method takes the data from each sample and with that data then calculates the z-score with respect to each species mean and then ranks the z-scores from lowest to highest.

$$\text{rank}\left(\frac{x-\mu_1}{\sigma_1},\ldots,\frac{x-\mu_k}{\sigma_k}\right)$$

This ranking is done for each trait and then totaled. The species with the lowest total is then selected. This is able to be done using up to k traits.
Concerns & Improvements:
There have been no cases where a trait ties, but that may be a future concern. Additionaly, we rank each trait immediately after receiving its z-score. Another viable option is totalling all of the z-scores and then only ranking once at the end. In our results some traits may be better indicators and instead of rating orderly (1st,2nd,…) the rankings could be scaled, e.g., ranking petal length heavily since that is a strong indicator of species, especially for setosta species.

Results:
Using all four traits.

```
##    Species
## dec setosa versicolor virginica
##   1     50          0         0
##   2      0         46        11
##   3      0          4        39
```

From the table above, we can see **using all four traits results in 90% accuracy (15 incorrect).** From the earlier plots, the values with the largest difference in means appeared to be petal length and petal length and for that reason we try the same method with only those two traits.

Using two traits, petal length and petal width.

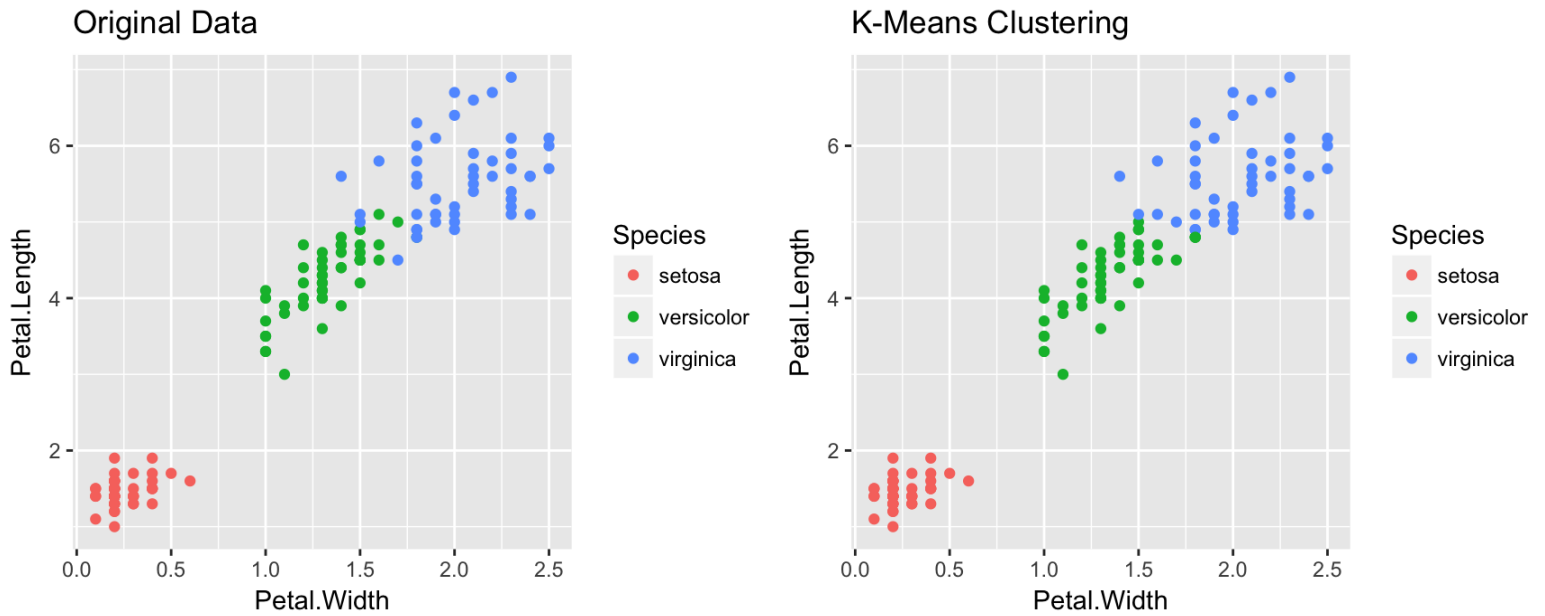
```
##    Species
## dec setosa versicolor virginica
##   1     50          0         0
##   2      0         49         7
##   3      0          1        43
```

Only utilizing those two traits allowed us to increase the **accuracy of our model up to 94.67% (8 incorrect).**

Classification by K-Means Clustering

Initially, we only cluster based on the petal width and the petal length. The reason being was those two traits appeared to have the largest difference in their means among species.
From this clustering we correctly classified 96% of the data (6 incorrect).

```
##
##     setosa versicolor virginica
##   1     50          0         0
##   2      0         48         4
##   3      0          2        46
```



We now test three traits: sepal width, petal length and petal width. **The result led to a less accurate result with 95.33% correct classification (7 incorrect features).**

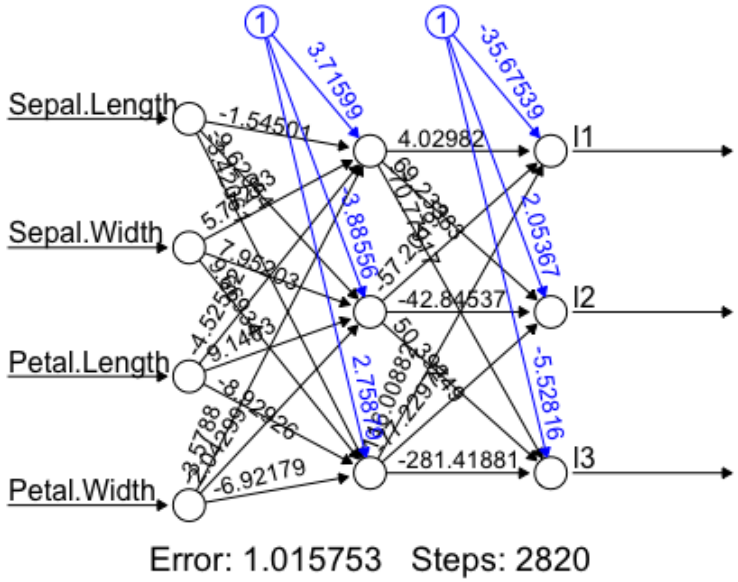
When all four traits: sepal length, sepal width, petal length and petal width, were used **the accuracy again decreased to 89.33% correct classification (16 incorrect)**

A valuable takeaway is, if it is possible to visually analyze the data prior, some helpful intuition can be gained on which characteristics to use for clustering. Sometimes less is more, parsimony principle!

Thoughts:
K-means is similar to the z-score classification and performed better without the prior knowledge of each species' mean value per trait. Difficulty may arise when the data is not so disjoint, but for situations like so the method performs effectively. Additionally, if the species type data was not included we could make an educated guess of which samples originated from each species, granted we know how many different species there are .

Nueral Networks

We fit our neural network with a (4,3,3) model with (3 being the hidden portion) as depicted below.



Results:

```
## hidden: 3    thresh: 0.01    rep: 1/1    steps:    1131  error: 1.84355  time: 0.21 secs
```

```
## [1] "Accuracy"
```

```
## [1] 0.9866666667
```

We were able to achieve 98.66% accuracy in correct classification. Out of all three models, nueral networks peformed the best and again did not need any prior knowlegde.

Concerns:
The model can be overfitting the data and for that reason we could apply a 10-fold cross validation only using 95% of the data and testing the remaining 5%

```
## [1] "Accuracy"
```

```
## [1] 0.9625
```

Of those 10 cross validation tests, the average accuracy was 96.25%.

Summary

In terms of accuracy, neural networks performed the best followed by k-cluster means and then z-score classification. For the z-score classification method some adjustments could be made that my significanlty improve the results but for a quick simple method the results were satisfactory. Other methods may be observed in the future for comparison.