**Question 4.2**

*The iris data set iris.txt contains 150 data points, each with four predictor variables and one categorical response. The predictors are the width and length of the sepal and petal of flowers and the response is the type of flower. The data is available from the R library datasets and can be accessed with iris once the library is loaded. It is also available at the UCI Machine Learning Repository (*[*https://archive.ics.uci.edu/ml/datasets/Iris*](https://archive.ics.uci.edu/ml/datasets/Iris) *). The response values are only given to see how well a specific method performed and should not be used to build the model.*

*Use the R function kmeans to cluster the points as well as possible. Report the best combination of predictors, your suggested value of k, and how well your best clustering predicts flower type.*

Here’s one possible solution. Please note that a good solution doesn’t have to try all of the possibilities in the code; they’re shown to help you learn, but they’re not necessary.

The R code in file solution 4.2.R shows clustering solutions for k=2,3,4,5 using all factors, for both unscaled and scaled data.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Unscaled data | | | | Scaled data | | | |
|  | Cluster | Setosa | Versi-color | Virgin-ica | Cluster | Setosa | Versi-color | Virgin-ica |
| k=2 | 1 | 50 | 3 | 0 | 1 | 50 | 0 | 0 |
| 2 | 0 | 47 | 50 | 2 | 0 | 50 | 50 |
| k=3 | 1 | 50 | 0 | 0 | 1 | 50 | 0 | 0 |
| 2 | 0 | 48 | 14 | 2 | 0 | 47 | 14 |
| 3 | 0 | 2 | 36 | 3 | 0 | 3 | 36 |
| k=4 | 1 | 50 | 0 | 0 | 1 | 50 | 0 | 0 |
| 2 | 0 | 27 | 1 | 2 | 0 | 27 | 2 |
| 3 | 0 | 0 | 32 | 3 | 0 | 0 | 29 |
| 4 | 0 | 23 | 17 | 4 | 0 | 23 | 19 |
| k=5 | 1 | 50 | 0 | 0 | 1 | 28 | 0 | 0 |
| 2 | 0 | 24 | 1 | 2 | 22 | 0 | 0 |
| 3 | 0 | 0 | 24 | 3 | 0 | 27 | 2 |
| 4 | 0 | 0 | 12 | 4 | 0 | 0 | 29 |
| 5 | 0 | 26 | 13 | 5 | 0 | 23 | 19 |

Table 1. Results using all factors

For k=2, the setosa species is almost perfectly in one cluster, and the other two species (versicolor and virginica) are in the other cluster. For k=3,4,5, setosa is a perfect cluster. When k=4,5 there’s a nice cluster of versicolor, a nice cluster or two of virginica, and a cluster of about 40 points that is mixed between the two. k=3 is a little more ambiguous – so even though there are 3 species, it turns out that k=4,5 work better.

The R code also shows clustering solutions for k=2,3,4,5 using only the Petal Length and Petal Width factors, for both unscaled and scaled data.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Unscaled data | | | | Scaled data | | | |
|  | Cluster | Setosa | Versi-color | Virgin-ica | Cluster | Setosa | Versi-color | Virgin-ica |
| k=2 | 1 | 50 | 1 | 0 | 1 | 50 | 0 | 0 |
| 2 | 0 | 49 | 50 | 2 | 0 | 50 | 50 |
| k=3 | 1 | 50 | 0 | 0 | 1 | 50 | 0 | 0 |
| 2 | 0 | 48 | 4 | 2 | 0 | 48 | 4 |
| 3 | 0 | 2 | 46 | 3 | 0 | 2 | 46 |
| k=4 | 1 | 50 | 0 | 0 | 1 | 50 | 0 | 0 |
| 2 | 0 | 26 | 0 | 2 | 0 | 42 | 0 |
| 3 | 0 | 0 | 35 | 3 | 0 | 0 | 27 |
| 4 | 0 | 24 | 15 | 4 | 0 | 8 | 23 |
| k=5 | 1 | 50 | 0 | 0 | 1 | 50 | 0 | 0 |
| 2 | 0 | 22 | 0 | 2 | 0 | 23 | 0 |
| 3 | 0 | 0 | 30 | 3 | 0 | 25 | 4 |
| 4 | 0 | 0 | 13 | 4 | 0 | 0 | 25 |
| 5 | 0 | 28 | 7 | 5 | 0 | 2 | 21 |

# Table 2. Results using only Petal Length and Petal Width factors

# Using only the Petal Length and Petal Width factors significantly improves the k=3 solution, and the k=5 solution. Notice that for k=4 especially, using scaled data is a big improvement over using unscaled data.

# The R code also introduces the ggplot2 library for plotting, just for your learning pleasure – it’s not required for the assignment.

# Of course, we can only create the tables above because we happen to know the correct species for each data point. Normally when we’re doing clustering, we don’t have that information. Instead, we can look at a measure like the total distance between points and their cluster centers in each clustering solution, as shown in the elbow diagram below for scaled data using only the petal factors.

Figure 1. Elbow diagram for scaled data using only petal factors.

Based on this figure, the 3-cluster solution might be the one we would recommend, since k=3 is where the improvements level out.