Homework Week 2, Question 4

2022-09-05

Question 4.1

Describe a situation or problem from your job, everyday life, current events, etc., for which a clustering model would be appropriate. List some (up to 5) predictors that you might use.

> My wife is a Realtor, and often times she must send out marketing campaigns to her lead database. This database includes information such as:
>
> 1. Year they last purchased
> 2. Annual Income
> 3. Size of current house
> 4. Number of family members
> 5. Location (Lattitude/ Longitude)
> 6. Age
>
> This information could all be used as predictors to segment her leads into different groups so she could tailor her messages to various groups. For example, her message to somone who just bought a new home this year, has a newborn baby, and lives in a 2b1b probably does not want to see the newest million dollar listing. They would be better off only getting a christmas card from my wife so her name sticks in their head for when they are ready to buy their new home.
>
> Conversely, she may find a cluster of leads who have been in their house for around 5 years, have $400k+ of annual income, and lives in a 5k sqft home. This person may be VERY interested in what their home is worth now and what homes in their price range are now on the market if they are looking for an upgrade.

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). 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.

> Per the below analysis, it appears the best value of K is k = 3, and the best predictors of species is *not* all of the measures, rather just petal length and width, which has a 96% accuracy.

Analysis

First, lets take a look at the data we have to work with.

```
head(iris)
```

```
##   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
```

Looks like we have 4 values that can be used to determine the Species. I'd like to look into what combination provides the best result. Should we use all four? Just the Sepal info? Or just the Petal info?

As a note, kmeans is an unsupervised learning method - we are going to assume we are building this model out to determine how many potential classifications there could be, rather than determining what classification the flower *is* given a set of data points. As such, we will only use the species information in determining which model is most accurate rather than using that information in our training data.

To scale, or not to scale?

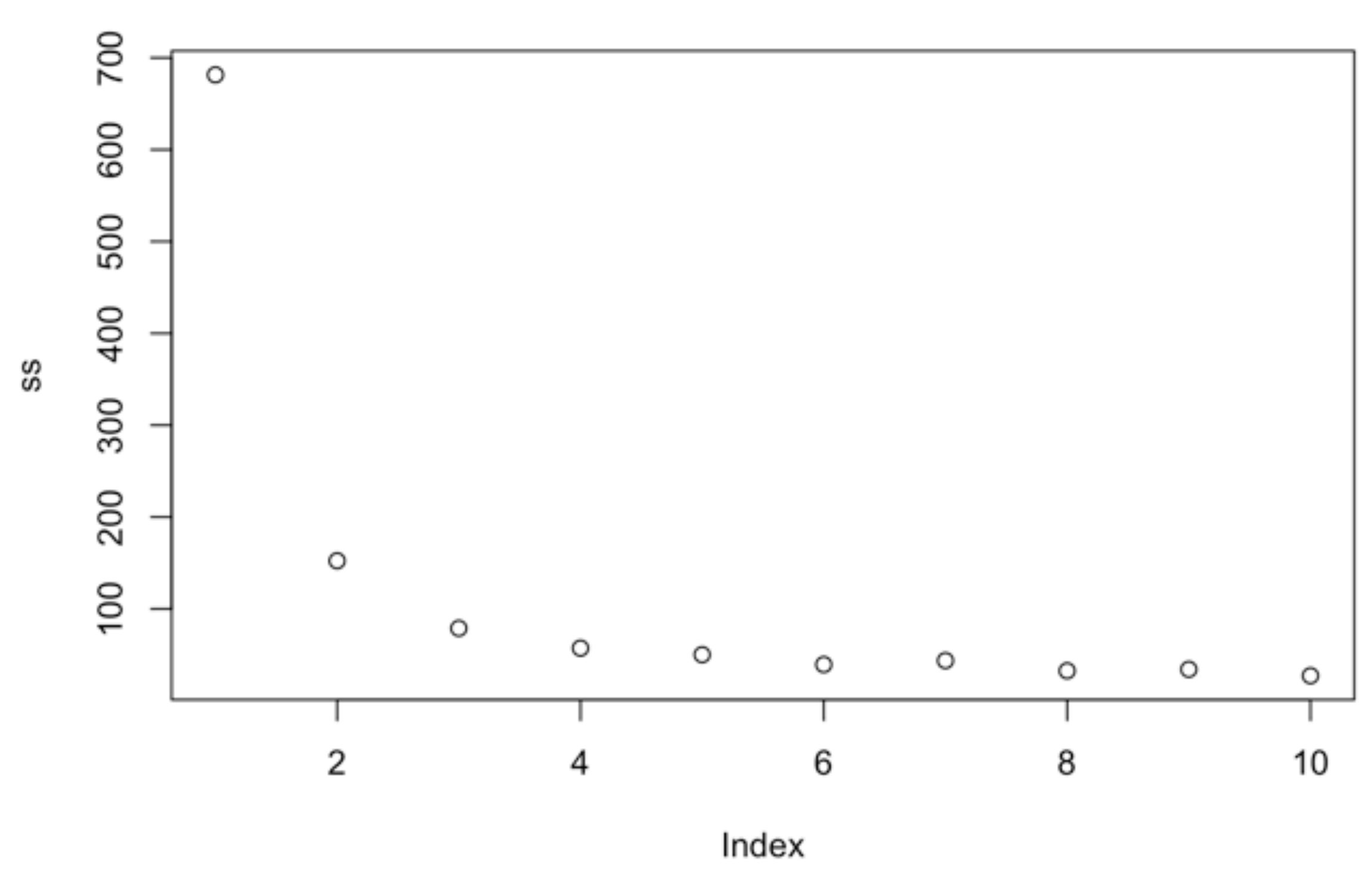
Typically, it is a good idea to scale data before using algorithms that are driven by distance, as the distance between predictors is not always 1:1. For example, if our data consisted of one measure in miles, and one measure in temperature, the delta of mile means something entirely different than a delta of 1 degree, so we scale the measures to make them equivalent.

However, with our iris data set, all measures are in the same measure of length, so scaling would not be appropriate, as we would lose the nuance of the distance measurements.

Determine how many classes to use

First, we want to see what the best value of k is to use. i.e. how many species should we separate iris flowers into? We will do this by using the "elbow" method, i.e. seeing what value of k gives the least sum of squares (distance between the points and the centroid) for each value of k, before our results begin to diminish.

```
k_list <- c(1:10)
ss <- c()
for (K in seq_along(k_list)){
  k_means.model.1 <- kmeans(iris[,1:4],K)
  ss[K] <- k_means.model.1$tot.withinss
}
plot(ss)
```



As we can see above, k = 3 seems to give us the most bang for our buck. Lets compare that to how many species there *actually* is in the data set.

```
length(unique(iris$Species))
```

```
## [1] 3
```

As it appears, there is only 3 different species of iris flowers! As such, we will use k= 3 for our models below.

Using Sepal and Petal Data

First, I'd like to see how accurate a model would be if we used all of the available data to predict what species a flower is given a new set of data. First we need to build the model.

```
set.seed(123)
k_means.model.1 <- kmeans(iris[,1:4],3)
k_means.model.1
```

```
## K-means clustering with 3 clusters of sizes 50, 62, 38
##
## Cluster means:
##   Sepal.Length Sepal.Width Petal.Length Petal.Width
## 1     5.006000    3.428000     1.462000    0.246000
## 2     5.901613    2.748387     4.393548    1.433871
## 3     6.850000    3.073684     5.742105    2.071053
##
## Clustering vector:
##   [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
##  [75] 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 2 3 3 3 3 2 3 3 3 3 3 3
## [112] 3 3 2 2 3 3 3 3 2 3 2 3 2 3 3 2 2 3 3 3 3 3 2 3 3 3 3 2 3 3 3 2 3 3 3 2 3
## [149] 3 2
##
## Within cluster sum of squares by cluster:
## [1] 15.15100 39.82097 23.87947
##  (between_SS / total_SS =  88.4 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

Now, let's compare our predictions against actual species classifications

```
table(k_means.model.1$cluster, iris$Species)
```

```
##
##     setosa versicolor virginica
##   1     50          0         0
##   2      0         48        14
##   3      0          2        36
```

It appears we have 125 data points (50 + 48 + 36) classified correctly, and 38 (36 + 2) classified incorrectly, giving us a 89% accuracy.

Sepal Length/Width

Now we will do the same for just sepal length/width.

```
k_means.model.2 <- kmeans(iris[,1:2],3)
table(k_means.model.2$cluster,iris$Species)
```

```
##
##     setosa versicolor virginica
##   1     50          0         0
##   2      0         38        15
##   3      0         12        35
```

Using Sepal length/width, we get 123 correct classifications, giving us an accuracy 82%

Petal Length/Width

```
k_means.model.3 <- kmeans(iris[,3:4],3)
table(k_means.model.3$cluster,iris$Species)
```

```
##
##     setosa versicolor virginica
##   1     50          0         0
##   2      0         48         6
##   3      0          2        44
```

Using just the petal length and width, we get 142 correct classifications and 8 incorrect, for an accuracy of 94.5%. Therefore, this is the best indicator of classification.