K-means clustering

Karen Mazidi

Getting set up

We are using the built-in iris data set.

```
library(datasets)
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
```

Now we try clustering with kmeans().

We are using just Petal.Length and Petal.Width for clustering. The number of clusters is set to 3 and the number of starts is 20.

```
set.seed(1234)
irisCluster <- kmeans(iris[, 3:4], 3, nstart=20)
irisCluster
```

```
## K-means clustering with 3 clusters of sizes 52, 48, 50
##
## Cluster means:
##   Petal.Length Petal.Width
## 1     4.269231    1.342308
## 2     5.595833    2.037500
## 3     1.462000    0.246000
##
## Clustering vector:
##   [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
##  [38] 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
##  [75] 1 1 1 2 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 1 2 2 2 2
## [112] 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2
## [149] 2 2
##
## Within cluster sum of squares by cluster:
## [1] 13.05769 16.29167  2.02200
##  (between_SS / total_SS =  94.3 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

Compare the clusters with the species. This is not usally something we can do in clustering because we normally don't have labels. We are usually clustering blind, not knowing the true grouping in the data.

```
table(irisCluster$cluster, iris$Species)
```

```
##
##     setosa versicolor virginica
```

```
##    1       0          48          4
##    2       0           2         46
##    3      50           0          0
```

Plot the clusters.

```
plot(iris$Petal.Length, iris$Petal.Width, pch=21, bg=c("red","green3","blue")
[unclass(irisCluster$cluster)], main="Iris Data")
```