MiniProj2

PetraGuy

15 January 2018

Within cluster sum of squares/between cluster sum of squares for the unscaled, semi-scaled and fully scaled data for ten repeats of kmeans.

```
##             [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## unscaled    0.14 0.14 0.14 0.14 0.14 0.14 0.14 0.14 0.14  0.14
## semi-scaled 0.34 0.34 0.34 0.34 0.34 0.34 0.34 0.34 0.34  0.34
## fullyscaled 0.42 0.42 0.42 0.42 0.42 0.42 0.42 0.42 0.42  0.42
```

The ratio is largest for the scaled data and does not decrease when the data is scaled, so an unscaled data set is preferable. The ratio is identical each time.

The accuracy over the ten repeats.

```
##             [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## unscaled    0.11 0.05 0.15 0.17 0.07 0.11 0.24 0.10 0.19  0.08
## semi-scaled 0.07 0.25 0.15 0.04 0.18 0.13 0.31 0.16 0.16  0.14
## fullyscaled 0.15 0.08 0.02 0.24 0.21 0.13 0.13 0.13 0.05  0.06
```

The accuracy is different for each repeat on all the data sets, suggesting that the algorithm is not successfully clustering the data.

The next tables show the percentage of each species correctly allocated to its cluster on each of the ten repeats

```
## [1] "unscaled"
```

```
##                2     3     4     5     6     7     8     9    10    11
## Anglica     2.50  3.12 18.75 26.25  3.12 18.75 26.25  0.00 20.62  0.00
## Cuneifolia  4.00  4.00 34.00  0.00  8.00  0.00 34.00 40.00  2.00  6.00
## Intermedia 31.58  0.00 21.05 26.32  0.00 26.32 21.05  0.00  0.00  0.00
## Leyana      6.25  6.25 10.42 31.25  6.25  6.25  6.25 29.17  6.25 16.67
## Minima     23.33 23.33  3.33  3.33 36.67  3.33  3.33 10.00  6.67  6.67
## Mougeotii  42.00  0.00  0.00  6.00  4.00  6.00 50.00  0.00 32.00 32.00
## Arranensis  0.00  4.35  0.00  0.00  0.00  0.00  0.00  0.00 73.91  0.00
```

```
## [1] "semi-scaled"
```

```
##                2     3     4     5     6     7     8     9    10    11
## Anglica     1.25 30.00  6.25  1.25 31.25  1.25 32.50 13.75  5.62 10.62
## Cuneifolia 22.00  0.00  8.00 26.00  0.00 42.00 38.00 30.00 10.00  0.00
## Intermedia 52.63  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00
## Leyana      0.00 43.75  0.00  2.08 20.83 45.83 45.83  0.00  0.00  0.00
## Minima      3.33 83.33  3.33  0.00  0.00  3.33  3.33  3.33  3.33  3.33
## Mougeotii   4.00  4.00 38.00  2.00 14.00  4.00 46.00 46.00 46.00 28.00
## Arranensis  0.00  0.00 95.65  0.00  0.00  0.00  0.00  0.00 91.30 95.65
```

```
## [1] "scaled"
```

```
##                2     3     4     5     6     7     8     9    10   11
## Anglica     0.00  1.88  0.00 30.63 49.38 18.12 28.75 28.75  0.00  0.0
## Cuneifolia 52.00 32.00 16.00  0.00  0.00  4.00  4.00  4.00 30.00  4.0
## Intermedia  0.00  5.26  0.00 94.74  0.00  0.00  0.00  0.00  0.00  0.0
## Leyana      6.25 20.83  2.08  2.08  2.08  6.25  4.17  2.08  0.00 12.5
## Minima     83.33  0.00  0.00  6.67  0.00  0.00  3.33  0.00  6.67  0.0
## Mougeotii   6.00  0.00  0.00 46.00  0.00 32.00  0.00  4.00  6.00 32.0
## Arranensis  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.0
```

The results again show that the algorithm is not consistently allocating species to the correct cluster. On some runs, is is very accurate for some species, but not necessarily for all the others. Then on other runs its is completely inaccurate.