chapter-clustering-validation

Ar

2022-07-19

Need for validation of clusters

Clustering algorithms are designed such that they come out with a given number of clusters even if the underlying data is devoid of any such clusters. We will see a criterion to assess the credibility of the clusters produced by any clustering algorithm.

Within-groups sum of squared distances (WSS):

$$WSS_k = \sum_{l=1}^{k} \sum_{x_i \in C_l} d^2(x_i, \overline{x_l})$$

where, k is the number of clusters and within the l-th cluster C_l, x_l is the centre of mass. We are interested in finding the *elbow* where there is a sudden drop in WSS_k as k is increased.

```
library("dplyr")
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
simdat = lapply(c(0,8), function(mx){

lapply(c(0,8), function(my){
tibble( x = rnorm(100, mean=mx, sd=2),
        y  = rnorm(100, mean=my, sd=2),
        class=paste(mx, my, sep=":"))

}) %>% bind_rows

}) %>% bind_rows
simdat
```

```
## # A tibble: 400 x 3
##          x       y class
##      <dbl>   <dbl> <chr>
##  1  1.87    0.0550 0:0
##  2  1.27    1.46   0:0
```

```
##  3  0.624  -1.02   0:0
##  4 -1.32   -0.413  0:0
##  5  0.804   0.406  0:0
##  6 -0.182   4.75   0:0
##  7 -2.32    1.34   0:0
##  8  0.582  -1.24   0:0
##  9  1.20    2.30   0:0
## 10 -0.0680 -1.57   0:0
## # ... with 390 more rows
```

```
library("tidyverse")
```

```
## Warning in system("timedatectl", intern = TRUE): running command 'timedatectl'
## had status 1
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.7     v stringr 1.4.0
## v tidyr   1.2.0     v forcats 0.5.1
## v readr   2.1.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
simdatxy = simdat[, c("x", "y")]
ggplot(simdat, aes(x=x, y=y, col=class))+geom_point()+coord_fixed()
```