Clustering

Aloksai Choudari, Alekhya Pinnamaneni

2022-10-09

Select a data set

Data set: Adults

Source: https://archive.ics.uci.edu/ml/datasets/Adult

Load in the data

```
adult <- read.csv("adult.data", header=FALSE)

# Adds columns names to the data table
colnames(adult) <- c('age', 'workclass', 'fnlwgt', 'education', 'education_num', 'marital_status', 'occu
```

Data Cleaning

```
# Changes the character columns to numeric columns
adult$workclass <- as.numeric(as.integer(factor(adult$workclass)))
adult$education <- as.numeric(as.integer(factor(adult$education)))
adult$marital_status <- as.numeric(as.integer(factor(adult$marital_status)))
adult$occupation <- as.numeric(as.integer(factor(adult$occupation)))
adult$relationship <- as.numeric(as.integer(factor(adult$relationship)))
adult$race <- as.numeric(as.integer(factor(adult$race)))
adult$sex <- as.numeric(as.integer(factor(adult$sex)))
adult$native_country <- as.numeric(as.integer(factor(adult$native_country)))
adult$predicted_salary_range <- as.numeric(as.integer(factor(adult$predicted_salary_range)))
```

kMeans clustering

```
library(datasets)
set.seed(1234)
cluster <- kmeans(adult[, 1:14], 2, nstart=20)
acc <- mean(cluster$cluster == adult$predicted_salary_range)
print(paste("accuracy = ", acc))
```

```
## [1] "accuracy =  0.38057799207641"
```

Model Based Clustering

```
library(mclust)
```

```
## Package 'mclust' version 5.4.10
## Type 'citation("mclust")' for citing this R package in publications.
```

```
cluster2 <- Mclust(adult)
acc2 <- mean(cluster2$classification == adult$predicted_salary_range)
print(paste("accuracy = ", acc2))
```

```
## [1] "accuracy =  0.376831178403612"
```

Analysis

The kMeans clustering algorithm had a slightly higher accuracy than the model based clustering algorithm. This difference can be explained by the iterative nature of the kMeans clustering algorithm. This algorithm iteratively assigns observations to the closest centroids and recalculates after each assignment. This results in a more accurate assignment of the data into clusters. The hierarchical clustering algorithm on the other hand has a bottom-up approach, which results in slightly less accurate results.