Data Mining in R

Pranshu Kumar

Analytics, CPS

Northeastern University

Table of Contents

[**Data Mining in R** 3](#_Toc31639325)

[Introduction 4](#_Toc31639326)

[Analysis 5](#_Toc31639327)

[Conclusion 11](#_Toc31639328)

[References 12](#_Toc31639329)

**Data Mining in R**

Using R to classify and cluster dataset to show how data mining methods can be used to classify and cluster data.

# Introduction

The project deals with working with practicing data mining techniques using R, and

The project deals with working with different packages and functions for building decision trees and produce a k-means cluster and a density based cluster.

## Analysis1

We started by loading the appropriate libraries needed for the project and analyed the ‘iris’ dataset.

The famous ‘iris’ dataset gives the ﻿measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica.

>install.packages("NbClust")

>install.packages("factoextra")

>?iris

> library(NbClust)

> library(factoextra)

> str(iris)

**## 'data.frame': 150 obs. of 5 variables:**

**## $ Sepal.Length: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...**

**## $ Sepal.Width : num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...**

**## $ Petal.Length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...**

**## $ Petal.Width : num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...**

**## $ Species : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...**

Now, we split the data into training and testing sets

>set.seed(1234)

> ind <- sample(2, nrow(iris), replace=T, prob=c(0.7, 0.3))

> iris.train <- iris[ind==1, ]

> iris.test <- iris[ind==2, ]

For plotting a decision tree, we need a the ‘party’ libarary.

> install.packages("party")

> library(party)

**## Loading required package: grid**

**## Loading required package: mvtnorm**

**## Loading required package: modeltools**

**## Loading required package: stats4**

**## Loading required package: strucchange**

**## Loading required package: zoo**

**##**

**## Attaching package: ‘zoo’**

**##**

**## The following objects are masked from ‘package:base’:**

**##**

**## as.Date, as.Date.numeric**

**##**

**## Loading required package: sandwich**

Now, we started building the decision tree.

>iris.form <- Species ~ Sepal.Length + Sepal.Width +

+ Petal.Length + Petal.Width

>iris.tree <- ctree(iris.form, data=iris.train)

> plot(iris.tree)

The decision tree looked like this –

Figure 1: Decision Tree



We predict on the test data and check the prediction result.

> prediction <- predict(iris.tree, newdata = iris.test)

> table(prediction, iris.test$Species)

**##**

**## prediction setosa versicolor virginica**

**## setosa 10 0 0**

**## versicolor 0 12 2**

**## virginica 0 0 14**

**K-Means Clustering**

We used the ‘fviz\_nbclust’ method from the ‘factoextra’ library, to find the optimal number of clusters for k-means clustering.

We firstly used the ‘elbow method’ of the ‘fviz\_nbclust’ on the ‘iris2’ dataframe.

>fviz\_nbclust(iris2, kmeans, method = "wss") +

+ geom\_vline(xintercept = 4, linetype = 2)+

+ labs(subtitle = "Elbow method")

**Figure 2: Result for Elbow method of fviz\_nbclust function**



We found the optimal number of clusters to be 4 according to Elbow method.

Similarly, tested the ‘**silhouette method**’.

>fviz\_nbclust(df, kmeans, method = "silhouette")+

+ labs(subtitle = "Silhouette method")

**Figure 3 – Result plot of silhouette method for the ‘fviz\_nbclust’ function**



The silhouette method suggests the optimal number of clusters to be **2**.

Then , we tested by **Gap Statistic method** –

> set.seed(123)

> fviz\_nbclust(iris2, kmeans, nstart = 25, method = "gap\_stat", nboot = 50)+

+ labs(subtitle = "Gap statistic method")

## Clustering k = 1,2,..., K.max (= 10): .. done

## Bootstrapping, b = 1,2,..., B (= 50) [one "." per sample]:

## .................................................. 50

**Figure 4 – Result plot for the Gap Statistic method of the fviz\_nbclust function**



The gap statistic method suggests the number of clusters as **6**.

Considering the observations , we decided to choose an average of **3 clusters** for the k-means clustering on the iris data.

We produced a k-means cluster on the same iris dataset.

> set.seed(8953)

> iris2 <- iris

# removed the class IDs for clustering

>iris2$Species <- NULL

> iris.kmeans <- kmeans(iris2, 3)

> table(iris$Species, iris.kmeans$cluster)

**##**

**## 1 2 3**

**## setosa 50 0 0**

**## versicolor 0 48 2**

**## virginica 0 14 36**

The table in the output shows the different species of the iris and the points assigned to each species based on clustering.

We plotted the clusters along with their centers.

> plot(iris2[c("Sepal.Length", "Sepal.Width")], col=iris.kmeans$cluster)

> points(iris.kmeans$centers[, c("Sepal.Length", "Sepal.Width")],

+ col=1:3, pch="\*", cex=5)



The density cluster shows figure shows the three different clusters formed from the iris data that represent the different types of the iris species.

**Density Based Clustering**

We produced a density based cluster.

>install.packages("fpc")

>library(fpc)

>iris2 <- iris[-5]

>ds <- dbscan(iris2, eps = 0.42, MinPts = 5)

> table(ds$cluster, iris$Species)

**##**

**## setosa versicolor virginica**

**## 0 2 10 17**

**## 1 48 0 0**

**## 2 0 37 0**

**## 3 0 3 33**

> plotcluster(iris2, ds$cluster)

**Figure 6 – Plot for Density based clustering**



The plot above shows the density based clustering with numbers assigned to each cluster with respect to density.

Conclusion

After successful completion of the project this week, we were successfully able to analyse the iris dataset. We successfully used the functions ‘sample()’, ‘ctree()’, ‘predict()’, provided by the ‘factoextra’ and ‘NbClust’ packages. We successfully built a decision tree and plotted it. Also, we found the appropriate number of clusters and produced a k-means cluster for the iris data, and further produced a density-based cluster. This all has led to successful applications of data mining techniques in R.

**References**

1. Introduction to Data Mining with R. (n.d.). Retrieved from http://www.rdatamining.com/docs/introduction-to-data-mining-with-r
2. Godfrey, K., Kassambara, Romero, J., Kassambara, Kumar, V., Kassambara, … G., G. (n.d.). Determining The Optimal Number Of Clusters: 3 Must Know Methods. Retrieved from https://www.datanovia.com/en/lessons/determining-the-optimal-number-of-clusters-3-must-know-methods/