Yuyi Zhang

May 04, 2019

Spring 2019 CPS Quarter Term A

Instructor: Steward Huang

Assignment 4: Data Mining

ALY 6015\_Intermediate Analytics

# **Introduction**

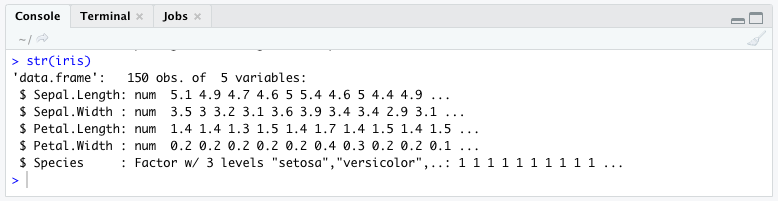
In this report, we are going to go through a data mining basic models applying in data example Iris. The dataset contains 150 observers and 5 variables. In the test, we are going to go through classification test, and clustering test by K-means clustering and density-based clustering.

# **Analysis**

## **Classification**

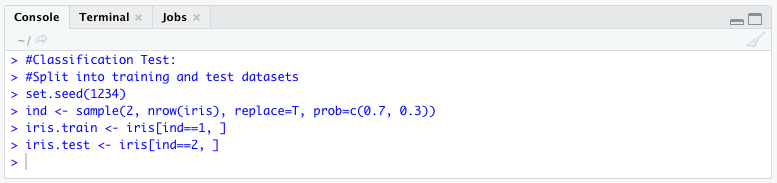
### **Data preparation**

In this phase, we are going use dataset “iris” and get to see the output of the snippet.



*Figure 1*. Data preparation

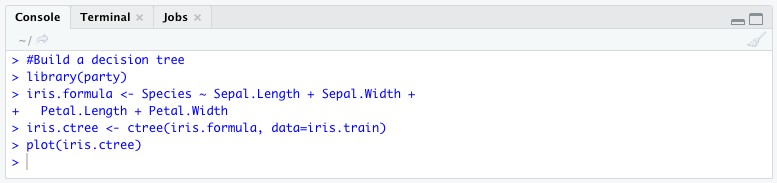
Then get the training data and test data split. In this case, we setup 70% of iris as training set, and the rest of them as testing set.



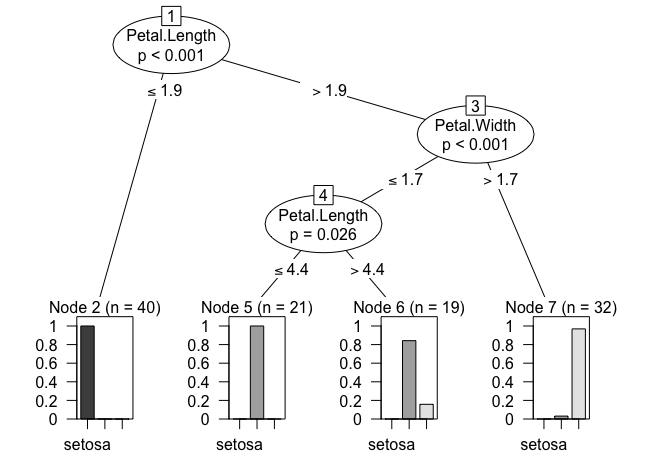
*Figure 2*. Split data

### **Build a decision tree**

We use training data to build up conditional inference trees to test the explanations of species with Sepal and Petal. Then plot the trees.



*Figure 3*. Build a decision tree

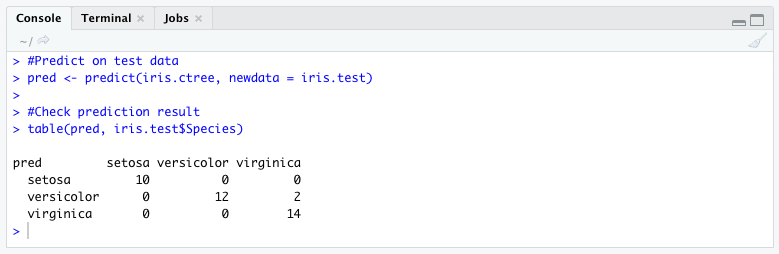


*Figure 4*. Conditional inference trees

The first split in Figure 4 separates Petal.length by greater than 1.9. Then the second split separates width and third, length again. As we can see that all of the nodes have pretty small p-value, so the species are well explained by petal length and width.

### **Prediction**

To see this table, the first column listed all three species. Every row represent the actual numbers and each column represent the test result. In setosa row, all the test results are equals to the actual results, as well as virginica row. The versicolor row had 2 items out of 14 results are not accurate. In general this is a pretty accurate and acceptable.



*Figure 5*. Prediction table

## **Clustering**

### **K-means Clustering**

For k-means clustering, we first removed the species name column and divided 150 items into 3 clusters. In the result table we could see that setosa and versicolor were decided quite accurate, but not the virginica.



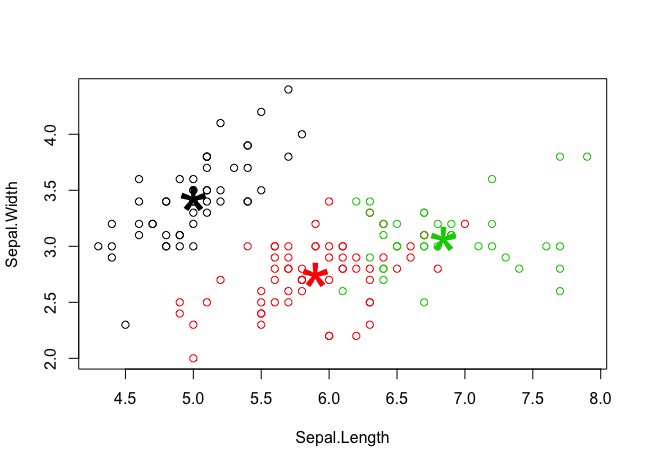
*Figure 6*. k-means Clustering

### **Plot clusters and their centers**

Plots three clusters based on sepal into three different colors. As we can see that the black cluster has been accurately divided. There were some overlaps between red and green spots.



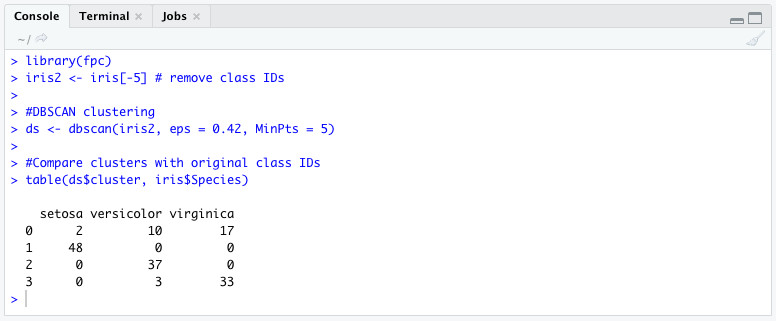
*Figure 7*. Plot clusters



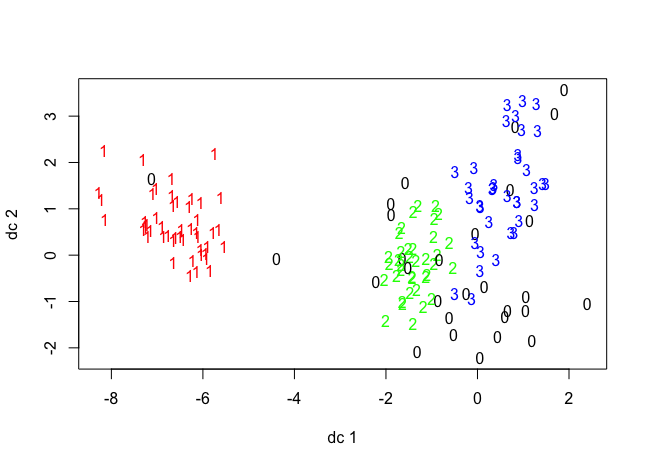
*Figure 8*. Cluster plots

### **Density-based Clustering**

In DBSCAN, we set up the max radius of the neighborhood as 0.42, and the min number of points in the Eps-neighborhood of a point is 5. Then we compare the results with original species. As the table showed, 29 out of 150 items in this dataset are noises.



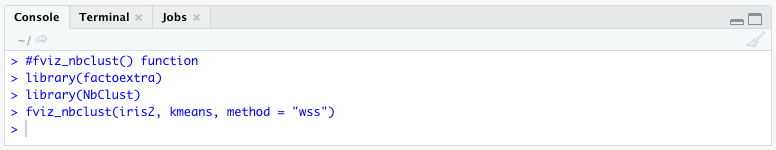
*Figure 8*. Density based clustering



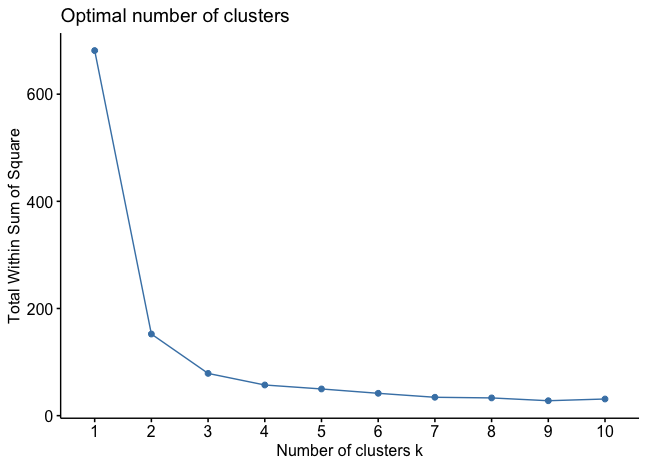
*Figure 9*. DBSCAN plots

### **Appendix: fviz\_nbclust() function**

Install related package then apply it to k-means clustering.



*Figure 10*. fviz.nbclust() function



*Figure 11*. optimal number of clusters

# **Conclusion**

Both k-means and density-based clustering are flat clustering algorithm. In this case, iris is a simple and straight dataset. Both classification and clustering told us that setosa could be better explained by Sepal size. Versicolor and virginica need some other various to be recognized.

Reference

1. Maindonald, J. H. (2008). *Using R for Data Analysis and Graphics.* Retrieved from <https://cran.r-project.org/doc/contrib/usingR.pdf>
2. Zhao, Y. (2015). *Introduction to Data Mining with R*. Retrieved from <http://www.rdatamining.com/docs/introduction-to-data-mining-with-r>