Capstone Project

CSE Training

Objective

This is a capstone project for CSE Training workshop, Introduction to R. In this project, we are going to ask you to conduct some analysis using a built-in data set in R. Specifically, you will use the famous data set, iris, which 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. You need to answer the following questions.

1. Build a predictive model (a decision tree) for the species of flowers using the variables in the data set: sepal length and width, petal length and width.
2. Suppose we do not have labels for the observations. Use K-Means method to perform clustering on the data set. Then compare the result with the real labels.

Useful hints and codes will be given throughout the following instructions. But there are many ways to achieve the above goals in R. You are highly encouraged to think of other (more efficient) ways of conducting the analysis in R. Your project should at least include these following parts:

- Reading and exploring data
- Building decision tree
- Clustering

Eventually, organize your project into a R markdown file and make it reproducible.

Functions

Functions that may be useful:

- Read and explore data: data(), str(), names(), summary(), boxplot(), hist()
- Decision Tree: ctree() (from package *party*), set.seed(), sample() (for dividing the data set into training and test sets) (HINT: Read the help files carefully)
- Clustering: kmeans()

Reading and Exploring Data

We start the analysis by reading the data and take a quick look at it.

```
data(iris);

names(iris);
```

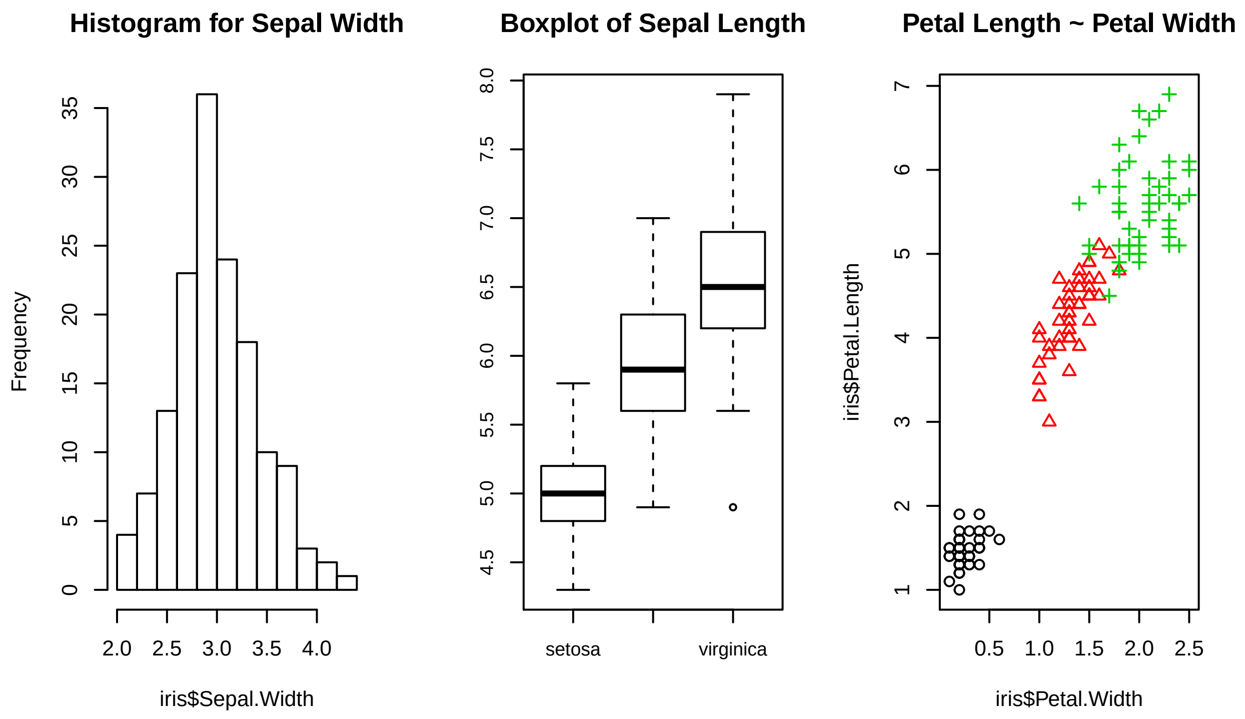
```
## [1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"
## [5] "Species"
```

```
str(iris);# Sometimes summary(iris) is also very helpful
```

```
## '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 ...
```

In order to get a more intuitive understanding of the data, we plot some exploratory graphs below.

```
par(mfrow = c(1,3));
hist(iris$Sepal.Width, main = "Histogram for Sepal Width");
boxplot(data = iris, Sepal.Length~Species,
        cex.axis = 0.85, main="Boxplot of Sepal Length");
plot(iris$Petal.Width, iris$Petal.Length,
     col=iris$Species, pch=as.numeric(iris$Species),
     main="Petal Length ~ Petal Width");
```



Build a Predictive Model

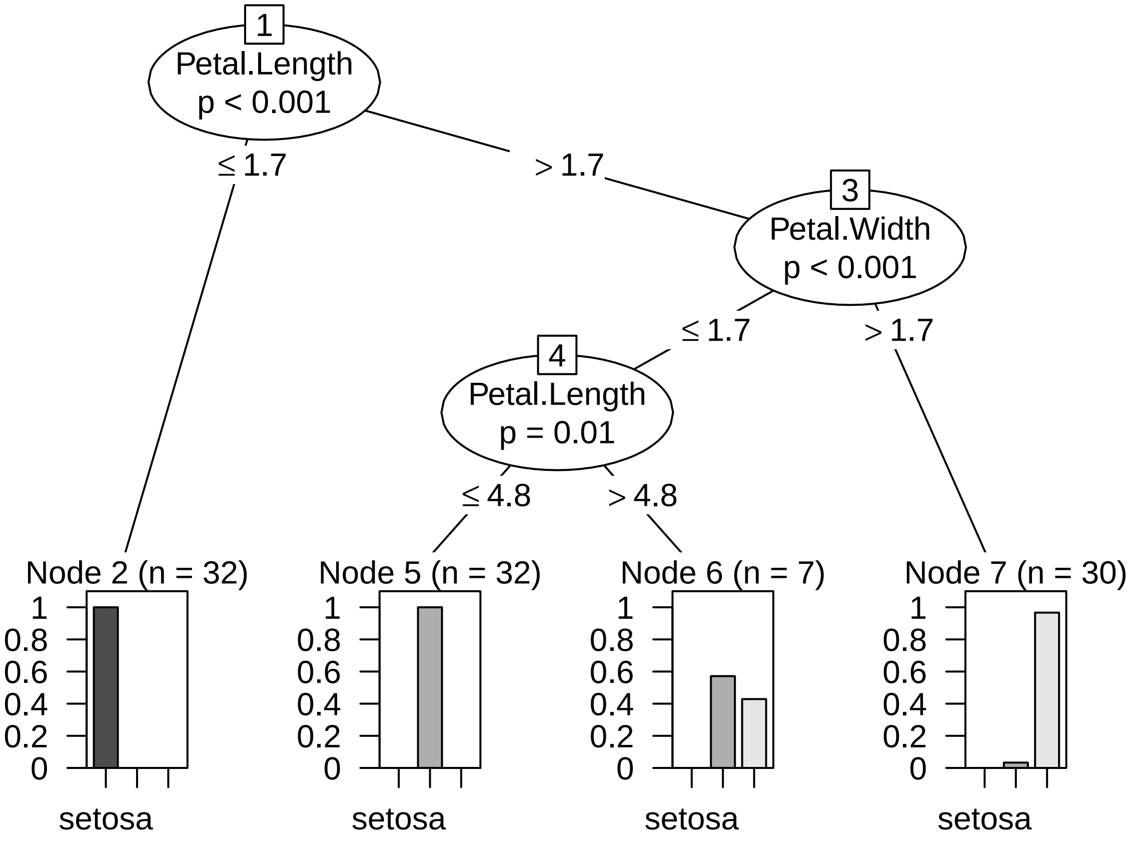
```
# Install package party first
library(party);

set.seed(3333)
index <- sample(2, nrow(iris), prob=c(0.7, 0.3), replace=TRUE)
train <- iris[index==1,]
test <- iris[index==2,]

predict_iris <- ctree(data=train, Species ~.);
print(predict_iris);
```

```
##
##   Conditional inference tree with 4 terminal nodes
##
## Response:  Species
## Inputs:  Sepal.Length, Sepal.Width, Petal.Length, Petal.Width
## Number of observations:  101
##
## 1) Petal.Length <= 1.7; criterion = 1, statistic = 93.618
##   2)*  weights = 32
## 1) Petal.Length > 1.7
##   3) Petal.Width <= 1.7; criterion = 1, statistic = 44.409
##     4) Petal.Length <= 4.8; criterion = 0.99, statistic = 9.088
##       5)*  weights = 32
##     4) Petal.Length > 4.8
##       6)*  weights = 7
##   3) Petal.Width > 1.7
##     7)*  weights = 30
```

```
plot(predict_iris);
```



```
# Do the prediction on test set and compare results with real labels
prediction <- predict(predict_iris, newdata=test);
table(prediction, test$Species);
```

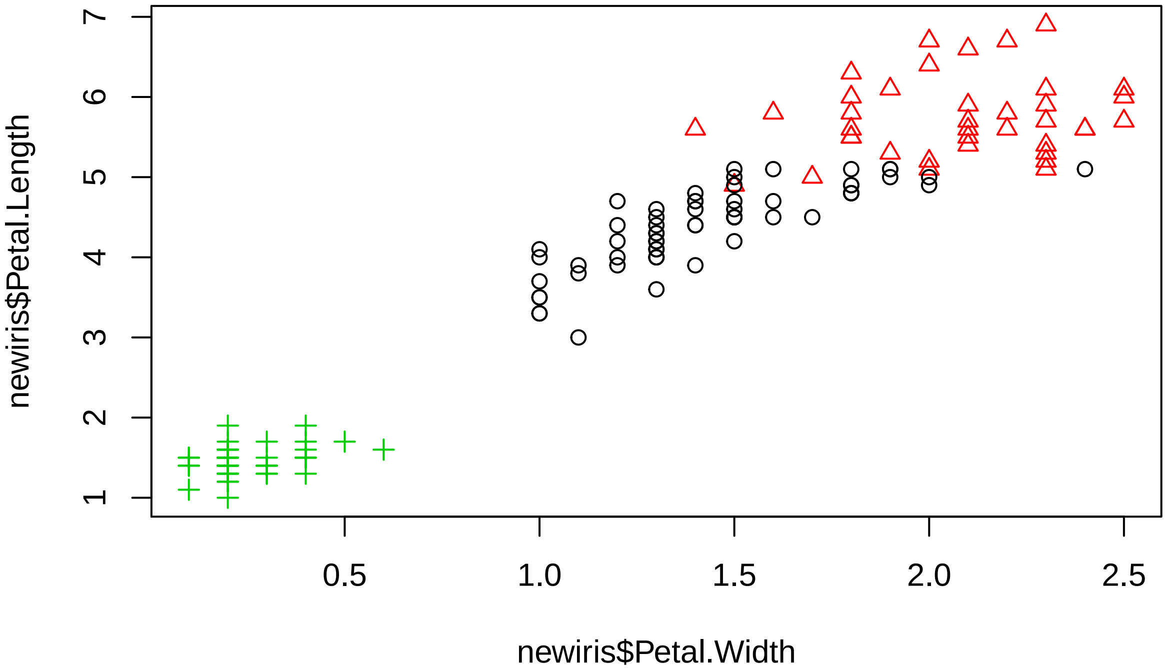
```
##
## prediction   setosa versicolor virginica
##   setosa         16          0         0
##   versicolor      2         13         2
##   virginica       0          0        16
```

Clustering

In this step, we move the labels away, and perform a K-Means clustering on the data set.

```
num_of_clusters <- 3;

newiris <- iris[,1:4];
kmean_iris <- kmeans(newiris, num_of_clusters);
plot(newiris$Petal.Width, newiris$Petal.Length,
     col=kmean_iris$cluster, pch=kmean_iris$cluster);
```



```
table(kmean_iris$cluster, iris$Species);
```

```
##
##     setosa versicolor virginica
##   1      0         48        14
##   2      0          2        36
##   3     50          0         0
```