Iris data analysis

Vaishnavi Bihare

5/1/2020

ANALYSIS USING CARET

Partioning the data into test and train data

validation\_index <- createDataPartition(i$Species, p=.80, list=FALSE)  
validation <- i[-validation\_index,]  
i <- i[validation\_index,]  
dim(i)

## [1] 120 5

dim(validation\_index)

## [1] 120 1

summary(validation\_index)

## Resample1   
## Min. : 1.00   
## 1st Qu.: 38.75   
## Median : 76.50   
## Mean : 76.02   
## 3rd Qu.:112.25   
## Max. :150.00

head(validation\_index)

## Resample1  
## [1,] 1  
## [2,] 2  
## [3,] 3  
## [4,] 4  
## [5,] 5  
## [6,] 7

head(validation)

## Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
## 6 5.4 3.9 1.7 0.4 setosa  
## 9 4.4 2.9 1.4 0.2 setosa  
## 11 5.4 3.7 1.5 0.2 setosa  
## 18 5.1 3.5 1.4 0.3 setosa  
## 21 5.4 3.4 1.7 0.2 setosa  
## 28 5.2 3.5 1.5 0.2 setosa

set.seed(150)  
sapply(i,class)

## Sepal.Length Sepal.Width Petal.Length Petal.Width Species   
## "numeric" "numeric" "numeric" "numeric" "factor"

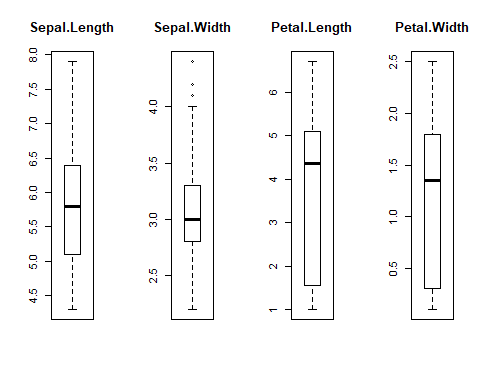
percentage <- prop.table(table(i$Species))\*100  
percentage

##   
## setosa versicolor virginica   
## 33.33333 33.33333 33.33333

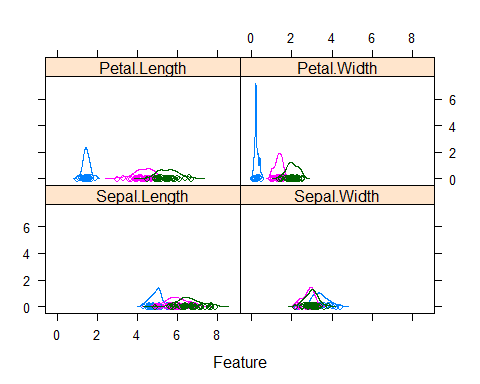
cbind(freq=table(i$Species), percentage=percentage)

## freq percentage  
## setosa 40 33.33333  
## versicolor 40 33.33333  
## virginica 40 33.33333

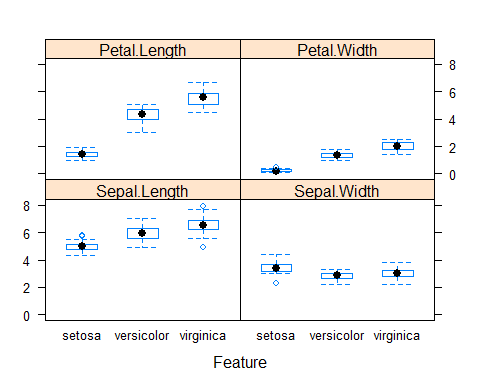
x <- i[,1:4]  
y <- i[,5]  
par(mfrow=c(1,4))  
for(a in 1:4) {  
 boxplot(x[,a], main=names(iris)[a])  
}



featurePlot(x=x,y=y,plot="density")



featurePlot(x=x,y=y,plot="box")



control <- trainControl(method = "cv", number=5)  
metric="Accuracy"

Fitting models

set.seed(100)  
fitlda <- train(Species~. ,data=i, method="lda", metric=metric, trcontrol=control)  
  
fitqda <- train(Species~. ,data=i, method="qda", metric=metric, trcontrol=control)  
#install.packages("rpart")  
#fitrpart <- train(Species~. ,data=i, method="rpart", metric=metric, trcontrol=control)  
results <- resamples(list(lda=fitlda, qda=fitqda))  
summary(results)

##   
## Call:  
## summary.resamples(object = results)  
##   
## Models: lda, qda   
## Number of resamples: 25   
##   
## Accuracy   
## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's  
## lda 0.9318182 0.9565217 0.975000 0.9673668 0.9772727 1 0  
## qda 0.9047619 0.9555556 0.974359 0.9660319 0.9795918 1 0  
##   
## Kappa   
## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's  
## lda 0.8963893 0.9332344 0.9621570 0.9504846 0.9651623 1 0  
## qda 0.8541667 0.9318182 0.9610778 0.9483000 0.9692982 1 0

print(fitlda)

## Linear Discriminant Analysis   
##   
## 120 samples  
## 4 predictor  
## 3 classes: 'setosa', 'versicolor', 'virginica'   
##   
## No pre-processing  
## Resampling: Bootstrapped (25 reps)   
## Summary of sample sizes: 120, 120, 120, 120, 120, 120, ...   
## Resampling results:  
##   
## Accuracy Kappa   
## 0.9673668 0.9504846

Prediction

prediction <- predict(fitlda,validation)  
confusionMatrix(prediction, validation$Species)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction setosa versicolor virginica  
## setosa 10 0 0  
## versicolor 0 10 0  
## virginica 0 0 10  
##   
## Overall Statistics  
##   
## Accuracy : 1   
## 95% CI : (0.8843, 1)  
## No Information Rate : 0.3333   
## P-Value [Acc > NIR] : 4.857e-15   
##   
## Kappa : 1   
##   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: setosa Class: versicolor Class: virginica  
## Sensitivity 1.0000 1.0000 1.0000  
## Specificity 1.0000 1.0000 1.0000  
## Pos Pred Value 1.0000 1.0000 1.0000  
## Neg Pred Value 1.0000 1.0000 1.0000  
## Prevalence 0.3333 0.3333 0.3333  
## Detection Rate 0.3333 0.3333 0.3333  
## Detection Prevalence 0.3333 0.3333 0.3333  
## Balanced Accuracy 1.0000 1.0000 1.0000