Assignment38

John Bute

2024-11-19

R Markdown

3. Construct and evaluate naïve Bayes classifiers for the Wine and for the 2011 Gapminder dataset.

```
library(e1071)
Wine <- read.csv("C:/Users/johnb/Desktop/Machine Learning/data/wine.csv",
stringsAsFactors = TRUE)

set.seed(1)
Wine$Class = as.factor(Wine$Class)

index <- sample(1:nrow(Wine), 0.7 * nrow(Wine))
train_data <- Wine[index, ]
test_data <- Wine[-index, ]

nb_model <- naiveBayes(Class ~ ., data = train_data)

summary(nb_model)
```

```
##           Length Class  Mode
## apriori    3     table  numeric
## tables    13     -none- list
## levels     3     -none- character
## isnumeric 13     -none- logical
## call       4     -none- call
```

```
nb_preds <- predict(nb_model, newdata = test_data, type = "class")
conf_matrix_nb <- table(Actual = test_data$Class, Predicted = nb_preds)
print(conf_matrix_nb)
```

```
##       Predicted
## Actual  1  2  3
##      1 20  0  0
##      2  0 21  1
##      3  0  0 12
```

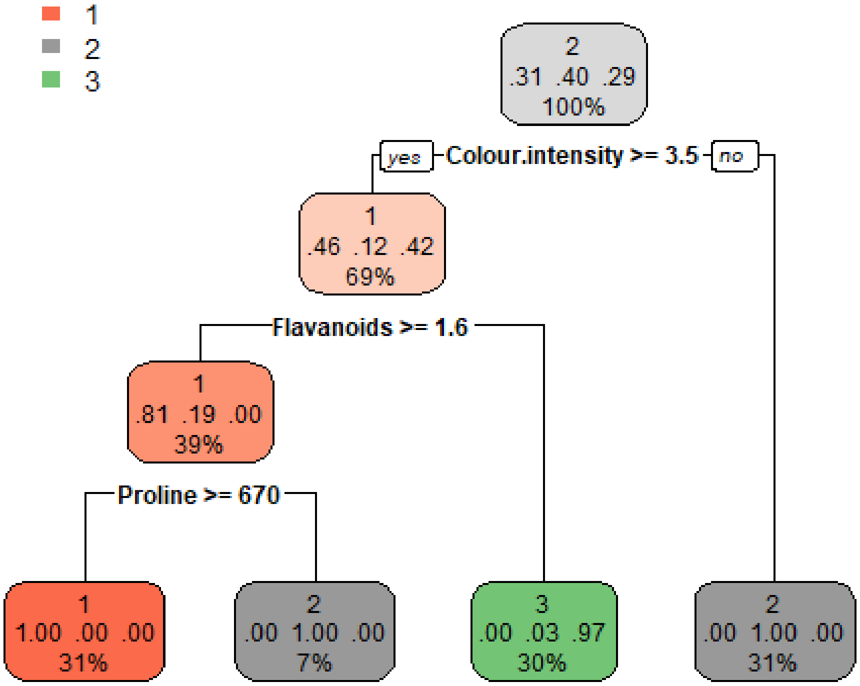
```
test_error <- 1 - sum(diag(conf_matrix_nb)) / sum(conf_matrix_nb)
cat("Test Error Rate:", test_error, "\n")
```

```
## Test Error Rate: 0.01851852
```

This model performs incredibly well in classifying all three wine classes across all metrics. It can predict classes 1 and 3 100% accurately, with a small mistep for class 2.

4. Construct and evaluate CART models for the Wine and for the Wisconsin Breast Cancer datasets.

```r
install.packages("rpart.plot")

## Installing package into 'C:/Users/johnb/AppData/Local/R/win-library/4.4'
## (as 'lib' is unspecified)

## package 'rpart.plot' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\johnb\AppData\Local\Temp\Rtmpcp8EXT\downloaded_packages

library(rpart)
library(rpart.plot)

## Warning: package 'rpart.plot' was built under R version 4.4.2

set.seed(1)
cart_model <- rpart(Class ~ ., data = train_data, method = "class")
rpart.plot(cart_model)
```



```r
summary(cart_model)

## Call:
## rpart(formula = Class ~ ., data = train_data, method = "class")
##   n= 124
##
##          CP nsplit  rel error    xerror       xstd
## 1 0.4333333      0 1.00000000 1.0000000 0.07258662
## 2 0.1200000      2 0.13333333 0.3200000 0.05865885
## 3 0.0100000      3 0.01333333 0.2933333 0.05672040
##
## Variable importance
##       Flavanoids    Total.phenols              Hue      OD280.OD315
##               13               12               11               11
## Colour.intensity       Malic.acid          Alcohol          Proline
##               10               10                9                8
##  Proanthocyanins        Magnesium              Ash
##                8                6                2
##
## Node number 1: 124 observations,    complexity param=0.4333333
##   predicted class=2  expected loss=0.6048387  P(node) =1
##     class counts:    39    49    36
##    probabilities: 0.315 0.395 0.290
##   left son=2 (85 obs) right son=3 (39 obs)
##   Primary splits:
##       Colour.intensity < 3.46  to the right, improve=31.23700, (0 missing)
##       Proline          < 755   to the right, improve=30.79193, (0 missing)
##       OD280.OD315      < 2.055 to the right, improve=29.85117, (0 missing)
##       Flavanoids       < 1.41  to the right, improve=29.28991, (0 missing)
##       Alcohol          < 12.78 to the right, improve=28.01821, (0 missing)
##   Surrogate splits:
##       Alcohol   < 12.35 to the right, agree=0.879, adj=0.615, (0 split)
##       Proline   < 411   to the right, agree=0.782, adj=0.308, (0 split)
##       Ash       < 2.02  to the right, agree=0.758, adj=0.231, (0 split)
##       Magnesium < 88.5  to the right, agree=0.750, adj=0.205, (0 split)
##       Hue       < 1.265 to the left,  agree=0.742, adj=0.179, (0 split)
##
## Node number 2: 85 observations,    complexity param=0.4333333
##   predicted class=1  expected loss=0.5411765  P(node) =0.6854839
##     class counts:    39    10    36
##    probabilities: 0.459 0.118 0.424
##   left son=4 (48 obs) right son=5 (37 obs)
##   Primary splits:
##       Flavanoids    < 1.58  to the right, improve=34.11141, (0 missing)
##       OD280.OD315   < 2.56  to the right, improve=33.27433, (0 missing)
##       Total.phenols < 2.335 to the right, improve=33.03823, (0 missing)
##       Hue           < 0.86  to the right, improve=27.03043, (0 missing)
##       Proline       < 755   to the right, improve=25.13086, (0 missing)
##   Surrogate splits:
##       OD280.OD315     < 2.385 to the right, agree=0.976, adj=0.946, (0
split)
##       Total.phenols   < 2.335 to the right, agree=0.953, adj=0.892, (0
split)
##       Hue             < 0.815 to the right, agree=0.918, adj=0.811, (0
split)
##       Proanthocyanins < 1.59  to the right, agree=0.871, adj=0.703, (0
split)
```

```
##       Malic.acid      < 2.18  to the left,  agree=0.859, adj=0.676, (0
split)
##
## Node number 3: 39 observations
##   predicted class=2  expected loss=0  P(node) =0.3145161
##     class counts:     0    39     0
##    probabilities: 0.000 1.000 0.000
##
## Node number 4: 48 observations,    complexity param=0.12
##   predicted class=1  expected loss=0.1875  P(node) =0.3870968
##     class counts:    39     9     0
##    probabilities: 0.812 0.187 0.000
##   left son=8 (39 obs) right son=9 (9 obs)
##   Primary splits:
##       Proline    < 670   to the right, improve=14.625000, (0 missing)
##       Magnesium  < 88.5  to the right, improve=10.820120, (0 missing)
##       Alcohol    < 12.99 to the right, improve= 4.548345, (0 missing)
##       Malic.acid < 1.465 to the right, improve= 4.548345, (0 missing)
##       Ash        < 2.245 to the right, improve= 3.125000, (0 missing)
##   Surrogate splits:
##       Magnesium     < 88.5  to the right, agree=0.958, adj=0.778, (0
split)
##       Alcohol       < 12.64 to the right, agree=0.917, adj=0.556, (0
split)
##       Malic.acid    < 1.3   to the right, agree=0.917, adj=0.556, (0
split)
##       Total.phenols < 2.23  to the right, agree=0.896, adj=0.444, (0
split)
##       Flavanoids    < 2.095 to the right, agree=0.896, adj=0.444, (0
split)
##
## Node number 5: 37 observations
##   predicted class=3  expected loss=0.02702703  P(node) =0.2983871
##     class counts:     0     1    36
##    probabilities: 0.000 0.027 0.973
##
## Node number 8: 39 observations
##   predicted class=1  expected loss=0  P(node) =0.3145161
##     class counts:    39     0     0
##    probabilities: 1.000 0.000 0.000
##
## Node number 9: 9 observations
##   predicted class=2  expected loss=0  P(node) =0.07258065
##     class counts:     0     9     0
##    probabilities: 0.000 1.000 0.000
```

```
cart_preds <- predict(cart_model, newdata = test_data, type = "class")
conf_matrix_cart <- table(Actual = test_data$Class, Predicted = cart_preds)
print(conf_matrix_cart)
```

```
##       Predicted
## Actual  1  2  3
##      1 20  0  0
##      2  2 18  2
##      3  0  0 12
```

```
test_error <- 1 - sum(diag(conf_matrix_cart)) / sum(conf_matrix_cart)
cat("Test Error Rate:", test_error, "\n")
```

```
## Test Error Rate: 0.07407407
```

Similarly, the cart model performs well, as it is strong in identifying class 1 and 3 wines, and overall, has an error rate of 7% only. Furthermore, the tree structure provides interpretable splits based on Flavanoids, Total.phenols, and hue.