Keosotra Veng
Stat 6620

Homework 5

Step 1: Step 1 - collecting data

. We got the data from:
`http://www.sci.csueastbay.edu/~esuess/classes/Statistics_6620/Presentations/ml10/redwines.csv`

Step 2: Exploring and preparing the data ----

- Load the data into R

```
redwine <-
read.csv("http://www.sci.csueastbay.edu/~esuess/classes/Statistics_6620/Prese
ntations/ml10/redwines.csv")
```

- Examine the redwine data

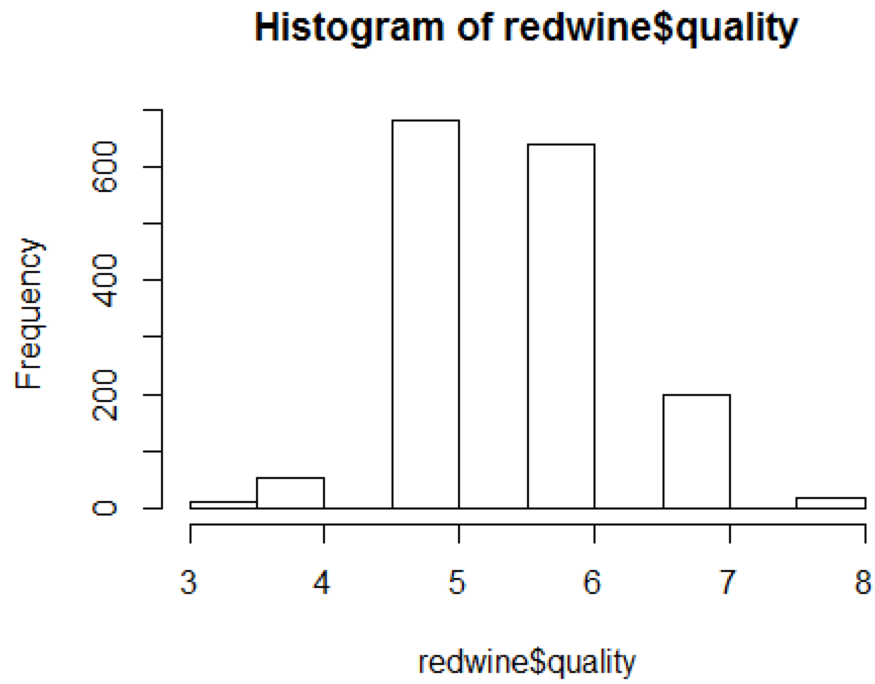
```
str(redwine)

## 'data.frame':    1599 obs. of  12 variables:
##  $ fixed.acidity       : num  6.5 9.1 6.9 7.3 12.5 5.4 10.4 7.9 7.3 9.5
...
##  $ volatile.acidity    : num  0.9 0.22 0.52 0.59 0.28 0.74 0.28 0.4 0.39
0.37 ...
##  $ citric.acid         : num  0 0.24 0.25 0.26 0.54 0.09 0.54 0.3 0.31
0.52 ...
##  $ residual.sugar      : num  1.6 2.1 2.6 2 2.3 1.7 2.7 1.8 2.4 2 ...
##  $ chlorides           : num  0.052 0.078 0.081 0.08 0.082 0.089 0.105
0.157 0.074 0.088 ...
##  $ free.sulfur.dioxide : num  9 1 10 17 12 16 5 2 9 12 ...
##  $ total.sulfur.dioxide: num  17 28 37 104 29 26 19 45 46 51 ...
##  $ density             : num  0.995 0.999 0.997 0.996 1 ...
##  $ pH                  : num  3.5 3.41 3.46 3.28 3.11 3.67 3.25 3.31 3.41
3.29 ...
##  $ sulphates           : num  0.63 0.87 0.5 0.52 1.36 0.56 0.63 0.91 0.54
0.58 ...
##  $ alcohol             : num  10.9 10.3 11 9.9 9.8 11.6 9.5 9.5 9.4 11.1
...
##  $ quality             : int  6 6 5 5 7 6 5 6 6 6 ...
```

- The distribution of quality ratings

```
hist(redwine$quality)
```

Keosotra Veng
Stat 6620



- Summary statistics of the redwine data

```
summary(redwine)

##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.01200   Min.   : 1.00        Min.   :  6.00
##  1st Qu.:0.07000   1st Qu.: 7.00        1st Qu.: 22.00
##  Median :0.07900   Median :14.00        Median : 38.00
##  Mean   :0.08747   Mean   :15.87        Mean   : 46.47
##  3rd Qu.:0.09000   3rd Qu.:21.00        3rd Qu.: 62.00
##  Max.   :0.61100   Max.   :72.00        Max.   :289.00
##     density             pH          sulphates         alcohol
##  Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   : 8.40
##  1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50
##  Median :0.9968   Median :3.310   Median :0.6200   Median :10.20
##  Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42
##  3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10
##  Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90
##     quality
##  Min.   :3.000
##  1st Qu.:5.000
##  Median :6.000
##  Mean   :5.636
##  3rd Qu.:6.000
##  Max.   :8.000

redwine_train <- redwine[1:1300, ]
redwine_test <- redwine[1301:1599, ]
```

Step 3: Training a model on the data ----

- regression tree using rpart

```
library(rpart)
m.rpart <- rpart(quality ~ ., data = redwine_train)
```

- Get basic information about the tree.

```
m.rpart

## n= 1300
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 1300 832.67310 5.642308
##    2) alcohol< 11.45 1066 560.36490 5.488743
##      4) sulphates< 0.645 640 260.36090 5.292188
##        8) volatile.acidity>=0.925 29  21.17241 4.551724 *
##        9) volatile.acidity< 0.925 611 222.53360 5.327332
##         18) alcohol< 9.975 358  96.70391 5.206704 *
##         19) alcohol>=9.975 253 113.24900 5.498024
##           38) free.sulfur.dioxide< 7.5 78  38.71795 5.205128 *
##           39) free.sulfur.dioxide>=7.5 175  64.85714 5.628571 *
##      5) sulphates>=0.645 426 238.13150 5.784038
##       10) alcohol< 9.95 170  78.35294 5.470588
##         20) free.sulfur.dioxide>=22.5 43  14.97674 5.023256 *
##         21) free.sulfur.dioxide< 22.5 127  51.85827 5.622047
##           42) fixed.acidity< 11.35 112  35.96429 5.517857 *
##           43) fixed.acidity>=11.35 15   5.60000 6.400000 *
##       11) alcohol>=9.95 256 131.98440 5.992188
##         22) volatile.acidity>=0.405 142  60.23239 5.781690 *
##         23) volatile.acidity< 0.405 114  57.62281 6.254386 *
##    3) alcohol>=11.45 234 132.64960 6.341880
##      6) sulphates< 0.635 103  55.96117 6.019417
##       12) pH>=3.265 70  31.78571 5.785714 *
##       13) pH< 3.265 33  12.24242 6.515152 *
##      7) sulphates>=0.635 131  57.55725 6.595420 *
```

- get more detailed information about the tree

```
summary(m.rpart)
```

Keosotra Veng
Stat 6620

```
## Call:
## rpart(formula = quality ~ ., data = redwine_train)
##   n= 1300
## 
##            CP nsplit rel error    xerror       xstd
## 1  0.16772319      0 1.0000000 1.0015353 0.04219529
## 2  0.07430590      1 0.8322768 0.8650627 0.04112546
## 3  0.03337941      2 0.7579709 0.8117606 0.04002416
## 4  0.02297559      3 0.7245915 0.7653762 0.03835458
## 5  0.02000181      4 0.7016159 0.7616070 0.03779887
## 6  0.01696845      5 0.6816141 0.7431779 0.03584983
## 7  0.01510873      6 0.6646456 0.7421791 0.03605306
## 8  0.01433099      7 0.6495369 0.7416133 0.03602419
## 9  0.01383247      8 0.6352059 0.7416133 0.03602419
## 10 0.01236257      9 0.6213734 0.7329247 0.03532889
## 11 0.01161791     10 0.6090109 0.7299716 0.03541873
## 12 0.01000000     11 0.5973930 0.7284213 0.03535684
## 
## Variable importance
##              alcohol            sulphates     volatile.acidity 
##                   34                   16                   10 
##              density        fixed.acidity          citric.acid 
##                    9                    8                    6 
##                   pH  free.sulfur.dioxide            chlorides 
##                    5                    4                    4 
## total.sulfur.dioxide       residual.sugar 
##                    2                    1 
## 
## Node number 1: 1300 observations,    complexity param=0.1677232
##   mean=5.642308, MSE=0.6405178 
##   left son=2 (1066 obs) right son=3 (234 obs)
##   Primary splits:
##       alcohol          < 11.45    to the left,  improve=0.16772320, (0 
missing)
##       sulphates        < 0.645    to the left,  improve=0.12354880, (0 
missing)
##       volatile.acidity < 0.425    to the right, improve=0.10502730, (0 
missing)
##       citric.acid      < 0.295    to the left,  improve=0.06961917, (0 
missing)
##       density          < 0.99537  to the right, improve=0.06695037, (0 
missing)
##   Surrogate splits:
##       density          < 0.994185 to the right, agree=0.875, adj=0.303, (0 
split)
##       fixed.acidity    < 5.5      to the right, agree=0.834, adj=0.077, (0 
split)
##       chlorides        < 0.0525   to the right, agree=0.832, adj=0.068, (0 
split)
##       pH               < 3.695    to the left,  agree=0.827, adj=0.038, (0
split)
##       volatile.acidity < 0.14     to the right, agree=0.822, adj=0.013, (0
split)
##
## Node number 2: 1066 observations,    complexity param=0.0743059
##   mean=5.488743, MSE=0.5256707
##   left son=4 (640 obs) right son=5 (426 obs)
##   Primary splits:
##       sulphates            < 0.645    to the left,  improve=0.11041470, (0
missing)
##       volatile.acidity     < 0.405    to the right, improve=0.09152379, (0
missing)
##       alcohol              < 9.975    to the left,  improve=0.08614538, (0
missing)
##       citric.acid          < 0.295    to the left,  improve=0.04422221, (0
missing)
##       total.sulfur.dioxide < 83.5     to the right, improve=0.03978478, (0
missing)
##   Surrogate splits:
##       citric.acid      < 0.395    to the left,  agree=0.682, adj=0.204, (0
split)
##       volatile.acidity < 0.4175   to the right, agree=0.681, adj=0.202, (0
split)
##       fixed.acidity    < 10.35    to the left,  agree=0.660, adj=0.150, (0
split)
##       pH               < 3.075    to the right, agree=0.636, adj=0.089, (0
split)
##       alcohol          < 10.525   to the left,  agree=0.636, adj=0.089, (0
split)
##
## Node number 3: 234 observations,    complexity param=0.02297559
##   mean=6.34188, MSE=0.5668785
##   left son=6 (103 obs) right son=7 (131 obs)
##   Primary splits:
##       sulphates        < 0.635    to the left,  improve=0.14422330, (0
missing)
##       citric.acid      < 0.325    to the left,  improve=0.09028404, (0
missing)
##       fixed.acidity    < 7.75     to the left,  improve=0.08734630, (0
missing)
##       pH               < 3.375    to the right, improve=0.05988673, (0
missing)
##       volatile.acidity < 0.425    to the right, improve=0.05682990, (0
missing)
##   Surrogate splits:
##       fixed.acidity    < 7.45     to the left,  agree=0.697, adj=0.311, (0
split)
##       citric.acid      < 0.285    to the left,  agree=0.684, adj=0.282, (0
split)
##       density          < 0.99411  to the left,  agree=0.667, adj=0.243, (0
split)
##       volatile.acidity < 0.5925   to the right, agree=0.628, adj=0.155, (0
split)
##       pH               < 3.415    to the right, agree=0.628, adj=0.155, (0
split)
##
## Node number 4: 640 observations,    complexity param=0.02000181
##   mean=5.292187, MSE=0.406814
##   left son=8 (29 obs) right son=9 (611 obs)
##   Primary splits:
##       volatile.acidity     < 0.925    to the right, improve=0.06396878, (0
missing)
##       sulphates            < 0.575    to the left,  improve=0.05770278, (0
missing)
##       alcohol              < 9.975    to the left,  improve=0.03445541, (0
missing)
##       pH                   < 3.425    to the right, improve=0.02079811, (0
missing)
##       total.sulfur.dioxide < 98.5     to the right, improve=0.01629491, (0
missing)
##   Surrogate splits:
##       total.sulfur.dioxide < 149.5    to the right, agree=0.956,
adj=0.034, (0 split)
##
## Node number 5: 426 observations,    complexity param=0.03337941
##   mean=5.784038, MSE=0.558994
##   left son=10 (170 obs) right son=11 (256 obs)
##   Primary splits:
##       alcohol              < 9.95     to the left,  improve=0.11671760, (0
missing)
##       volatile.acidity     < 0.405    to the right, improve=0.10992940, (0
missing)
##       chlorides            < 0.0965   to the right, improve=0.09323926, (0
missing)
##       total.sulfur.dioxide < 50.5     to the right, improve=0.09215330, (0
missing)
##       density              < 0.996225 to the right, improve=0.05193817, (0
missing)
##   Surrogate splits:
##       chlorides        < 0.1045   to the right, agree=0.678, adj=0.194, (0
split)
##       sulphates        < 0.975    to the right, agree=0.646, adj=0.112, (0
split)
##       volatile.acidity < 0.565    to the right, agree=0.636, adj=0.088, (0
split)
##       pH               < 3.045    to the left,  agree=0.631, adj=0.076, (0
split)
##       residual.sugar   < 1.85     to the left,  agree=0.629, adj=0.071, (0
split)
##
```

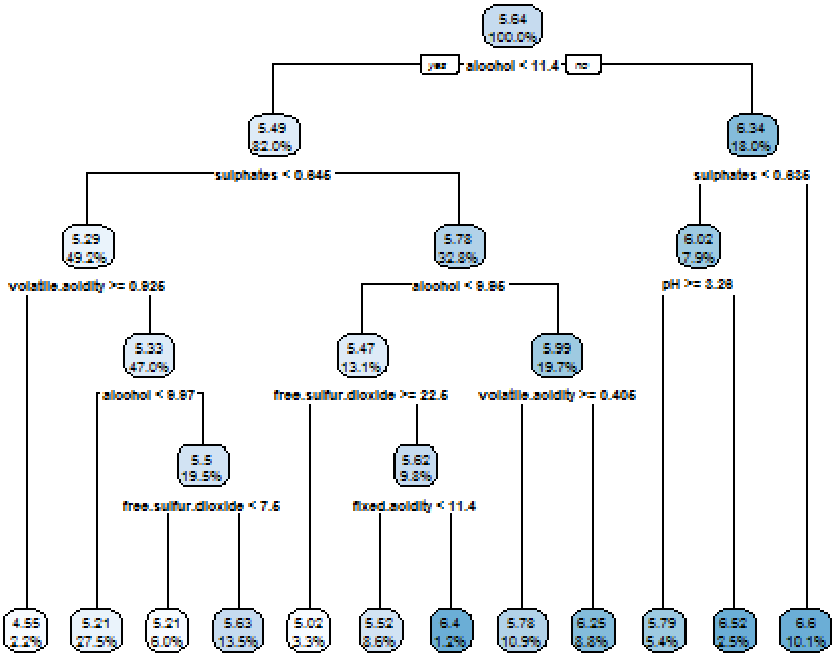
```
## Node number 6: 103 observations,    complexity param=0.01433099
##   mean=6.019417, MSE=0.5433123 
##   left son=12 (70 obs) right son=13 (33 obs)
##   Primary splits:
##       pH                   < 3.265    to the right, improve=0.2132376, (0 
missing)
##       citric.acid          < 0.445    to the left,  improve=0.1525071, (0 
missing)
##       volatile.acidity     < 0.495    to the right, improve=0.1354948, (0 
missing)
##       free.sulfur.dioxide  < 31.5     to the left,  improve=0.1332219, (0 
missing)
##       fixed.acidity        < 6.55     to the left,  improve=0.1044414, (0 
missing)
##   Surrogate splits:
##       citric.acid          < 0.335    to the left,  agree=0.874, adj=0.606, 
(0 split)
##       fixed.acidity        < 7.8      to the left,  agree=0.864, adj=0.576, 
(0 split)
##       volatile.acidity     < 0.385    to the right, agree=0.806, adj=0.394, 
(0 split)
##       chlorides            < 0.0995   to the left,  agree=0.748, adj=0.212, 
(0 split)
##       free.sulfur.dioxide  < 34       to the left,  agree=0.748, adj=0.212, 
(0 split)
## 
## Node number 7: 131 observations
##   mean=6.59542, MSE=0.4393683 
## 
## Node number 8: 29 observations
##   mean=4.551724, MSE=0.7300832 
## 
## Node number 9: 611 observations,    complexity param=0.01510873
##   mean=5.327332, MSE=0.364212 
##   left son=18 (358 obs) right son=19 (253 obs)
##   Primary splits:
##       alcohol               < 9.975    to the left,  improve=0.05653363, (0 
missing)
##       sulphates             < 0.575    to the left,  improve=0.05173810, (0 
missing)
##       volatile.acidity      < 0.6525   to the right, improve=0.03180631, (0 
missing)
##       total.sulfur.dioxide  < 98.5     to the right, improve=0.02422109, (0 
missing)
##       density               < 0.99569  to the right, improve=0.01771703, (0 
missing)
##   Surrogate splits:
##       density               < 0.995805 to the right, agree=0.678, 
adj=0.221, (0 split)
##       total.sulfur.dioxide  < 37.5     to the right, agree=0.640, 
adj=0.130, (0 split)
##       fixed.acidity        < 6.95     to the right, agree=0.622,
adj=0.087, (0 split)
##       chlorides            < 0.0685   to the right, agree=0.622,
adj=0.087, (0 split)
##       sulphates            < 0.595    to the left,  agree=0.615,
adj=0.071, (0 split)
##
## Node number 10: 170 observations,    complexity param=0.01383247
##   mean=5.470588, MSE=0.4608997
##   left son=20 (43 obs) right son=21 (127 obs)
##   Primary splits:
##       free.sulfur.dioxide  < 22.5     to the right, improve=0.14700060, (0
missing)
##       fixed.acidity        < 11.8     to the left,  improve=0.14369840, (0
missing)
##       volatile.acidity     < 0.3175   to the right, improve=0.12410440, (0
missing)
##       total.sulfur.dioxide < 46.5     to the right, improve=0.12406210, (0
missing)
##       chlorides            < 0.0955   to the right, improve=0.07724758, (0
missing)
##   Surrogate splits:
##       total.sulfur.dioxide < 66.5     to the right, agree=0.865,
adj=0.465, (0 split)
##       residual.sugar       < 3.25     to the right, agree=0.800,
adj=0.209, (0 split)
##       density              < 1.0009   to the right, agree=0.771,
adj=0.093, (0 split)
##       volatile.acidity     < 0.855    to the right, agree=0.765,
adj=0.070, (0 split)
##       sulphates            < 1.6      to the right, agree=0.753,
adj=0.023, (0 split)
##
## Node number 11: 256 observations,    complexity param=0.01696845
##   mean=5.992188, MSE=0.515564
##   left son=22 (142 obs) right son=23 (114 obs)
##   Primary splits:
##       volatile.acidity     < 0.405    to the right, improve=0.10705190, (0
missing)
##       total.sulfur.dioxide < 54.5     to the right, improve=0.09371862, (0
missing)
##       residual.sugar       < 3.8      to the right, improve=0.04513882, (0
missing)
##       chlorides            < 0.0975   to the right, improve=0.04398857, (0
missing)
##       pH                   < 3.48     to the right, improve=0.03639320, (0
missing)
##   Surrogate splits:
##       citric.acid    < 0.305    to the left,  agree=0.719, adj=0.368, (0
split)
##       sulphates      < 0.765    to the left,  agree=0.621, adj=0.149, (0
split)
##       chlorides      < 0.0675   to the right, agree=0.617, adj=0.140, (0
split)
##       residual.sugar < 1.85     to the right, agree=0.602, adj=0.105, (0
split)
##       fixed.acidity  < 7.55     to the left,  agree=0.590, adj=0.079, (0
split)
##
## Node number 12: 70 observations
##   mean=5.785714, MSE=0.4540816
##
## Node number 13: 33 observations
##   mean=6.515152, MSE=0.3709826
##
## Node number 18: 358 observations
##   mean=5.206704, MSE=0.2701227
##
## Node number 19: 253 observations,    complexity param=0.01161791
##   mean=5.498024, MSE=0.4476246
##   left son=38 (78 obs) right son=39 (175 obs)
##   Primary splits:
##       free.sulfur.dioxide  < 7.5      to the left,  improve=0.08542167, (0
missing)
##       total.sulfur.dioxide < 14.5     to the left,  improve=0.05140778, (0
missing)
##       sulphates            < 0.585    to the left,  improve=0.04450115, (0
missing)
##       volatile.acidity     < 0.655    to the right, improve=0.02790290, (0
missing)
##       pH                   < 3.405    to the right, improve=0.02470002, (0
missing)
##   Surrogate splits:
##       total.sulfur.dioxide < 16.5     to the left,  agree=0.881,
adj=0.615, (0 split)
##       alcohol              < 11.35    to the right, agree=0.723,
adj=0.103, (0 split)
##       pH                   < 3.55     to the right, agree=0.711,
adj=0.064, (0 split)
##       sulphates            < 0.45     to the left,  agree=0.711,
adj=0.064, (0 split)
##       chlorides            < 0.1445   to the right, agree=0.708,
adj=0.051, (0 split)
##
## Node number 20: 43 observations
##   mean=5.023256, MSE=0.3482964
##
## Node number 21: 127 observations,    complexity param=0.01236257
##   mean=5.622047, MSE=0.4083328
##   left son=42 (112 obs) right son=43 (15 obs)
##   Primary splits:
##       fixed.acidity    < 11.35    to the left,  improve=0.19850220, (0
missing)
##       density          < 0.99716  to the left,  improve=0.16814750, (0
missing)
##       volatile.acidity < 0.3175   to the right, improve=0.13160050, (0
missing)
##       alcohol          < 9.85     to the left,  improve=0.10955790, (0
missing)
##       pH               < 2.99     to the right, improve=0.09450066, (0
missing)
##   Surrogate splits:
##       volatile.acidity < 0.235    to the right, agree=0.898, adj=0.133, (0
split)
##       density          < 0.99965  to the left,  agree=0.898, adj=0.133, (0
split)
##       pH               < 2.89     to the right, agree=0.898, adj=0.133, (0
split)
##       citric.acid      < 0.71     to the left,  agree=0.890, adj=0.067, (0
split)
##
## Node number 22: 142 observations
##   mean=5.78169, MSE=0.4241718
##
## Node number 23: 114 observations
##   mean=6.254386, MSE=0.5054632
##
## Node number 38: 78 observations
##   mean=5.205128, MSE=0.496384
##
## Node number 39: 175 observations
##   mean=5.628571, MSE=0.3706122
##
## Node number 42: 112 observations
##   mean=5.517857, MSE=0.3211097
##
## Node number 43: 15 observations
##   mean=6.4, MSE=0.3733333
```

- use the rpart.plot package to create a visualization

```
library(rpart.plot)
```

- a basic decision tree diagram

```
rpart.plot(m.rpart, digits = 3)
```



Step 4: Evaluate model performance ----

- generate predictions for the testing dataset

```
p.rpart <- predict(m.rpart, redwine_test)
```

- compare the distribution of predicted values vs. actual values

```
summary(p.rpart)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.552   5.207   5.518   5.603   5.784   6.595

summary(redwine_test$quality)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   5.000   6.000   5.609   6.000   8.000
```

- compare the correlation. This shows that our prediction is not too bad with 0.61 correlation.

```
cor(p.rpart, redwine_test$quality)

## [1] 0.6092693
```

- function to calculate the mean absolute error

```
MAE <- function(actual, predicted) {
  mean(abs(actual - predicted))
}
```

- mean absolute error between predicted and actual values

```
MAE(p.rpart, redwine_test$quality)

## [1] 0.5447783
```

- mean absolute error between actual values and mean value

```
mean(redwine_train$quality) # result = 5.87

## [1] 5.642308

MAE(5.87, redwine_test$quality)

## [1] 0.7020401
```

Step 5: Improving model performance ----

- train a M5' Model Tree

```
library(RWeka)
m.m5p <- M5P(quality ~ ., data = redwine_train)
```

- display the tree

```
m.m5p

## M5 pruned model tree:
## (using smoothed linear models)
## LM1 (1300/81.02%)
##
## LM num: 1
## quality =
##  -0.9688 * volatile.acidity
##  - 2.0438 * chlorides
##  + 0.0054 * free.sulfur.dioxide
##  - 0.0033 * total.sulfur.dioxide
##  - 0.4868 * pH
##  + 0.874 * sulphates
##  + 0.2777 * alcohol
##  + 4.5326
##
## Number of Rules : 1
```

- get a summary of the model's performance.

```
summary(m.m5p)

##
## === Summary ===
##
## Correlation coefficient                  0.5862
## Mean absolute error                      0.5027
## Root mean squared error                  0.6484
## Relative absolute error                 74.2453 %
## Root relative squared error             81.02   %
## Total Number of Instances             1300
```

- generate predictions for the model

```
p.m5p <- predict(m.m5p, redwine_test)
```

- summary statistics about the predictions

```
summary(p.m5p)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   4.711   5.249   5.487   5.594   5.941   6.745
```

- correlation between the predicted and true values. Our correlation is 0.657 which is slightly higher than the previous model which was about 0.61.

```
cor(p.m5p, redwine_test$quality)

## [1] 0.6576933
```

- mean absolute error of predicted and true values
- (uses a custom function defined above)

```
MAE(redwine_test$quality, p.m5p)

## [1] 0.5029122
```