Janne Flinck

79428

I start by splitting the data into a training set and a testing set. The training set consists of 70% of randomly sampled data and the testing set consists of the remaining 30%.

I start the linear regression by removing the variables ”density” from the list of variables for the linear regression model, because of its high absolute correlation with residual sugar and alcohol. Since I am dealing with a regression problem, I am going to use the mean squared error (MSE) as a measure of how much our predictions are far away from the real data. My out-of-sample MSE is 0.544 and in-sample MSE is 0.567 with this model.

```
Call:
lm(formula = quality ~ . - density, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.8156 -0.4935 -0.0341  0.4643  2.7363 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           5.883414   0.013055 450.677  < 2e-16 ***
fixed.acidity        -0.042541   0.015319  -2.777 0.005517 ** 
volatile.acidity     -0.192080   0.013794 -13.925  < 2e-16 ***
citric.acid          -0.009656   0.014186  -0.681 0.496115    
residual.sugar        0.142881   0.015613   9.152  < 2e-16 ***
chlorides            -0.007904   0.013977  -0.566 0.571739    
free.sulfur.dioxide   0.070860   0.017067   4.152 3.38e-05 ***
total.sulfur.dioxide -0.028744   0.019300  -1.489 0.136486    
pH                    0.026640   0.015084   1.766 0.077466 .  
sulphates             0.045114   0.013350   3.379 0.000735 ***
alcohol               0.460309   0.016569  27.781  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7642 on 3418 degrees of freedom
Multiple R-squared:  0.2687,	Adjusted R-squared:  0.2666 
F-statistic: 125.6 on 10 and 3418 DF,  p-value: < 2.2e-16
```

Confidence intervals:

```
                      2.5 % 97.5 %
(Intercept)           5.858  5.909
fixed.acidity        -0.073 -0.013
volatile.acidity     -0.219 -0.165
citric.acid          -0.037  0.018
residual.sugar        0.112  0.173
chlorides            -0.035  0.019
free.sulfur.dioxide   0.037  0.104
total.sulfur.dioxide -0.067  0.009
pH                   -0.003  0.056
sulphates             0.019  0.071
alcohol               0.428  0.493
```

With the KNN regression, the k that minimizes the out-of-sample error rate is 8, with an error of 0.506. My in-sample MSE is 0.393.