STAT 6340 Statistical Machine Learning

Mini Project 2

Author:
Suarez Rodes, Randy

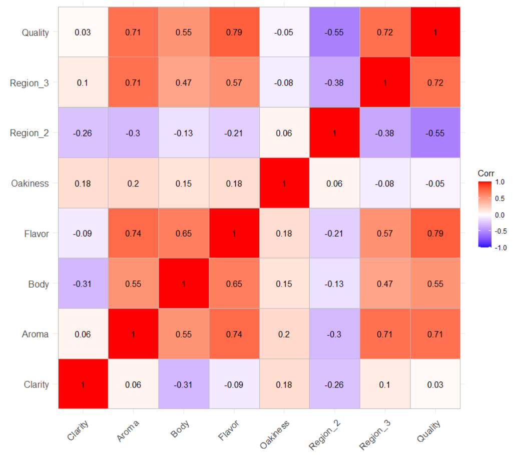
Supervisor:
Choudhary, Pankaj Ph.D.

February 21, 2021

1 Answers

Question 1.a

We first explore the correlations of our variables, both predictors and response on Figure 1. Visualizing the correlation matrix we can see that predictors Oakiness and Clarity have a very low correlation with Quality. We infer that these variables are not much relevant to predict Quality and are the most likely to be removed from future models. On the other hand predictors Aroma, Flavor and Region has a high correlation with Quality, making them good candidates for our models.
Another point to notice is that variable Body presents some correlation with Quality, although, it is not as high as other predictors, therefore we require further analysis to determine its relevance for the response variable. Finally, predictors Aroma, Body, Flavor and Region are highly correlated. This can cause overfitting issues, since the data would be over explained, so with further analysis one or some of this predictors might be dropped since they can be explained by the others.



	Clarity	Aroma	Body	Flavor	Oakiness	Region_2	Region_3	Quality
Quality	0.03	0.71	0.55	0.79	-0.05	-0.55	0.72	1
Region_3	0.1	0.71	0.47	0.57	-0.08	-0.38	1	0.72
Region_2	-0.26	-0.3	-0.13	-0.21	0.06	1	-0.38	-0.55
Oakiness	0.18	0.2	0.15	0.18	1	0.06	-0.08	-0.05
Flavor	-0.09	0.74	0.65	1	0.18	-0.21	0.57	0.79
Body	-0.31	0.55	1	0.65	0.15	-0.13	0.47	0.55
Aroma	0.06	1	0.55	0.74	0.2	-0.3	0.71	0.71
Clarity	1	0.06	-0.31	-0.09	0.18	-0.26	0.1	0.03

Figure 1: Correlation Matrix

We present the scatterplot of our variables on Figure 2. As expected from the correlation matrix we can notice that the predictors Aroma, Body and Flavor exhibit a positive linear relation with Quality.
On Figure 3 we present the Quality boxplots by Region. Pay attention to the clear distribution of Quality by Region (high on 3, medium on 1, low on 2). This fact evidences the importance of this predictor for determining the Quality of a wine.

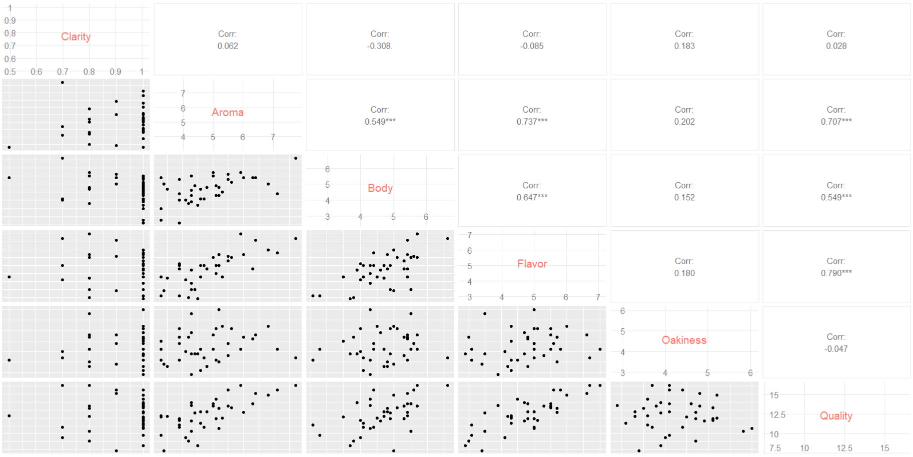


Figure 2: Scatter Plots

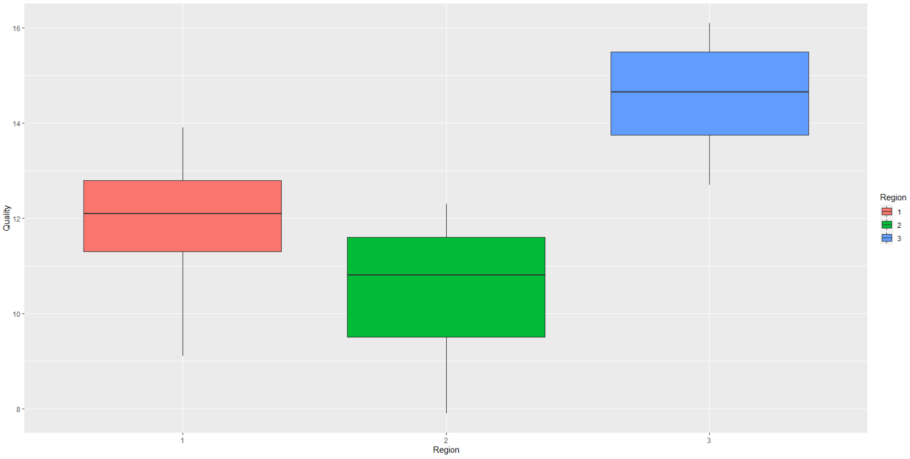


Figure 3: Quality by Region

Question 1.b

Question 1.c We fitted a simple linear regression model for every predictor. Taking a look to the t-test of every model we can affirm that there is a statistically significant association between the predictor and the response variable on all models except for the models with predictors Clarity (p-value = 0.865) and Oakiness (p-value = 0.779). This is consistent with our previous exploration. We can see on Figure 4 how there is a clear linear relation for predictors Flavor, Aroma and Body. Also we discussed on question 1.a how Region is relevant to predict Quality (Figure 3).

Question 1.d
From the summary of the full model we can reject the null hypothesis for the Flavor and Region predictors, meaning that this predictors are relevant in order to predict Quality.

Question 1.e
First we will proceed to drop one variable at a time in order of from the full model, going in the order from the highest p-value and rechecking hypothesis test result for every new model.
After this process we ended up with a the model Quality ~Flavor + Region.

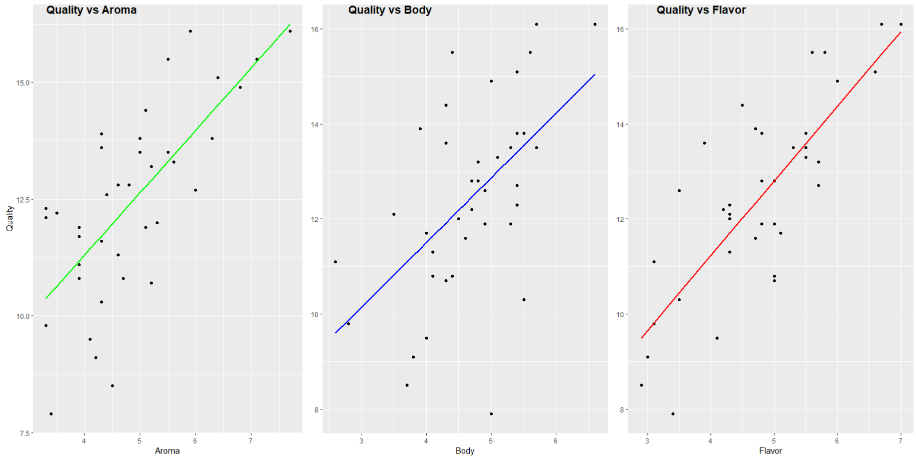


Figure 4: Quality by Region

Now we will proceed to explore interactions between these two predictors.

We will compare the model Quality ~Flavor + Region with the model Quality ~Flavor + Region + Flavor:Region. Carry on an anova test we obtained a p-value of 0.3378, meaning that we fail to reject the null hypothesis, this is all additional predictors are 0. Therefore the interactions between Flavor and Region are not meaningful for our model.

We can appreciate the three resulting regression lines of our model for each region on figure 4. Clearly our model captures the linear trend on each region

Also we present some diagnostics plots on figure 5 in order to check the models assumptions.
We can see in the Residuals vs Fitted plot how the points are scattered around zero and there is no a clear pattern, verifying that our errors have an approximate zero mean and constant variance. For the Q-Q Plot, observe how for an exception of a few points the errors are close to be normally distributed. Finally checking on the Residuals Time Series, there is not a clear dependence of the residuals, verifying the independence assumption.

Question 1.f

Final model:
Quality = Intercept + Flavor + Region2 + Region3

More precisely:
Quality = 7.0943 + 1.1155(Flavor) -1.5335($\mathbb{1}_{Region2}$) + 1.2234($\mathbb{1}_{Region3}$)

We can interpret our model as follows: Approximately every level change in Flavor will directly change the Quality by one level. Region 1 wine has base quality of 7, Region 3 offers an additional level (quality base around 8) and Region 2 a decrease of 1.5 (quality base approximately of 5.5).

Question 1.g

Our prediction of Quality for a wine from Region 1 with a mean Flavor is: 12.41371

We obtained a 95% confidence interval of [11.95152, 12.8759], this is that given an average Flavor wine and assuming it is from Region 1, 95% of intervals of this form will contain the true value of the wine Quality.

Also we obtained a 95% prediction interval of [10.53775, 14.28967], meaning that we can be 95% confident that the Quality of an average Flavor wine from Region 1 will lie in this range.

Question 2.a

On figure 6 we present several plots to explore how our predictors can help to predict the group of an applicant. We can observe in the graph of GMAT vs GPA, how the GPA predictor can separate the three groups. Moreover in the GPA boxplot we have a clear distribution of the GPA by group, evidence of how much this predictor contributes to identify the correct group. For the GMAT boxplot there is some confusion for applicants of group 2 and 3, but a clear distinction for applicants of group 1, meaning that GMAT also help us characterizes the possible group of an applicant.

Question 2.b

We present the confusion matrix for the best value of K(50) on Table 2. A quick look at the confusion matrix immediately shows that the misclassification is high. In particular we can quickly find the accuracy yielding 36% which is as bad as random guess. We can notice that many observations have been wrongly classified as class 2 (birds), only one observation from class 7 (horses) has been correctly classified and also only one for class 1 (automobile).

Table 1: Test Error Rate

kval	Error
50	0.64
100	0.67
200	0.67
300	0.66
400	0.67

Table 2: Confusion Matrix

		Actual									
		0	1	2	3	4	5	6	7	8	9
Predicted	0	2	1	0	0	0	0	1	0	1	0
	1	0	1	0	0	0	0	0	0	1	0
	2	2	1	3	4	6	5	1	2	2	3
	3	0	0	0	1	0	0	0	0	0	0
	4	0	0	1	4	5	0	2	2	0	3
	5	0	0	0	1	0	1	0	1	0	0
	6	1	2	1	1	0	1	8	0	2	0
	7	0	0	0	0	0	0	0	1	0	0
	8	1	2	0	0	1	2	0	1	12	3
	9	0	1	0	0	0	0	0	0	1	2

Question 2.c

It is not a good idea to use KNN for image classification since the KNN algorithm relies on a distance metric to find similarities. This is a problem for high-dimensional objects like images, since distances over high dimensional spaces can be very counter intuitive. Also, depending on the distance, some transformation might be require leading to possible loss of information. Another problem is that KNN considers all features equally important for computing the similarities of observations. This can lead to misclassification in many situations where the algorithm determines two objects are close using irrelevant features like color or size in many cases.

Question 2.d

Question 3.a

By definition: $MSE\{\hat{f}(x_0)\} = E[\hat{f}(x_0) - f(x_0)]^2$,
$Bias\{\hat{f}(x_0)\} = E[\hat{f}(x_0)] - f(x_0)$ and $var\{\hat{f}(x_0)\} = E(\hat{f}(x_0)^2 - E[f(x_0)])^2$

Then: $(Bias\{\hat{f}(x_0)\})^2 = E[\hat{f}(x_0)]^2 - 2E[\hat{f}(x_0)]f(x_0) + f(x_0)^2$

$var\{\hat{f}(x_0)\} = E([\hat{f}(x_0)^2 - E[\hat{f}(x_0)][\hat{f}(x_0)^2 - E[\hat{f}(x_0)]]) = E[\hat{f}(x_0)^2] - 2E[\hat{f}(x_0)]E[\hat{f}(x_0)] +$

$E[\hat{f}(x_0)]^2 = E[\hat{f}(x_0)^2] - E[\hat{f}(x_0)]^2$

Combining the Bias and Variance terms we obtain: $E[\hat{f}(x_0)]^2 - 2E[\hat{f}(x_0)]f(x_0) + f(x_0)^2 + E[\hat{f}(x_0)^2] - E[\hat{f}(x_0)]^2 = E[\hat{f}(x_0)^2] - 2E[\hat{f}(x_0)]f(x_0) + f(x_0)^2$

Also note that: $MSE\{\hat{f}(x_0)\} = E[\hat{f}(x_0) - f(x_0)]^2 = E[\hat{f}(x_0)^2] - 2E[\hat{f}(x_0)f(x_0)] + E[f(x_0)^2] = E[\hat{f}(x_0)^2] - 2E[\hat{f}(x_0)]f(x_0) + f(x_0)^2$
Since $f(x_0)$ is not a random variable and its expectation is equal to itself.

Therefore from both expressions we have that $MSE\{\hat{f}(x_0)\} = (Bias\{\hat{f}(x_0)\})^2 + var\{\hat{f}(x_0)\}$

Now by definition: $Y_0 = f(x_0) + \epsilon_0$, $\hat{Y}_0 = \hat{f}(x_0)$
and $\sigma^2 = E[\epsilon_0^2 - E(\epsilon_0)]^2 = E(\epsilon_0^2) - [E(\epsilon_0)]^2 = E(\epsilon_0^2)$, since $E(\epsilon_0) = 0$. Then:

$E(\hat{Y}_0 - Y_0)^2 = E[\hat{f}(x_0) - f(x_0) - \epsilon_0]^2 = E[\hat{f}(x_0) - f(x_0)]^2 - 2E\{[\hat{f}(x_0) - f(x_0)]\epsilon_0\} + E(\epsilon_0^2) =$

$MSE\{\hat{f}(x_0)\} - 2E[(\hat{f}(x_0) - f(x_0)]E[\epsilon_0] + \sigma^2 = MSE\{\hat{f}(x_0)\} + \sigma^2$ QED.
Where in the last two steps we have used the independence of ϵ_0 and the fact that $E(\epsilon_0) = 0$

Question 3.b

Question 3.c

Question 3.d

2 Code