

# STAT 6340 Statistical Machine Learning

## Mini Project 2

*Author:*

Suarez Rodes, Randy

*Supervisor:*

Choudhary, Pankaj Ph.D.

February 20, 2021



# 1 Answers

## Question 1.a

We first explore the correlations of our variables, both predictors and response. 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.

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, average on 1, low on 2). This fact evidences the importance of this predictor.

## 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 3 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).

## 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  $\text{Quality} \sim \text{Flavor} + \text{Region}$ . Now we will proceed to explore interactions between these two predictors.

We will compare the model  $\text{Quality} \sim \text{Flavor} + \text{Region}$  with the model  $\text{Quality} \sim \text{Flavor} + \text{Region} + \text{Flavor}:\text{Region}$  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.

## Question 1.f

Final model:

$\text{Quality} = \text{Intercept} + \text{Flavor} + \text{Region2} + \text{Region3}$

More precisely:

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

**Question 1.g****Question 2.a**

We have fitted KNN for the testing set for values of  $K = 50, 100, 200, 300$  and  $400$ . The error rates are presented on Table 1. Observe the high error rates for the different values of  $K$ . Concluding that the KNN classifier did not perform well on the testing set.

**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 2: Confusion Matrix

Table 1: Test Error Rate

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

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

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

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

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

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:

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

$$MSE\{\hat{f}(x_0)\} - 2E[(\hat{f}(x_0) - f(x_0))E[\epsilon_0]] + \sigma^2 = MSE\{\hat{f}(x_0)\} + \sigma^2 \text{ QED.}$$

Where in the last two steps we have used the independence of  $\epsilon_0$  and the fact that  $E(\epsilon_0) = 0$

**Question 3.b**

**Question 3.c**

**Question 3.d**

## 2 Code