Abstract

We build a model that predicts a wine's quality based on its chemical properties. We use a Bayesian linear regression as a model and can thus determine the posterior probability of a wine being of outstanding quality. Applying this model out of sample will show us that vodka might be the best wine of them all . . . or maybe the model should only be applied to Portuguese red wines.

1 Introduction

We want to predict the quality and taste of a red wine without time and cost intensive surveys. Instead we want to use the chemical attributes of a wine to predict its quality on a scale from 1 to 10. We are particularly interested in the probability that the wine scores at least a 7/10 as these exclusive wines might distributed through different channels with a more tailored pricing.

2 Data

We use a wine dataset made available by Cortez et al. (2009). It contains the chemical properties and quality of 1599 Portuguese wines. The data has the following columns:

1. fixed acidity
2. volatile acidity
3. citric acid
4. residual sugar
5. chlorides
6. free sulfur dioxide
7. total sulfur dioxide
8. density
9. pH
10. sulphates
11. alcohol
12. quality

This data obviously fits the problem we are trying to answer. The data is already well cleaned and has no missing values or obvious errors.

The data was also graphically analysed. As an example Figure 1 shows the correlation between the different columns of the red wine dataset. We can see that most variables have only a week linear correlation with the wine's quality. The quality is most strongly correlated with the alcohol content and the volatile acidity.

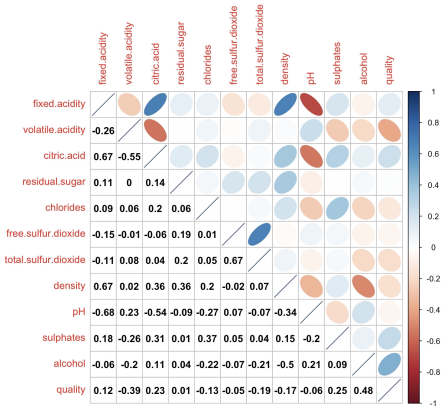

Figure 1: Correlation plot of the data

3 Model

To predict the quality scores of different wines a regression model is needed and a linear model seems a good first guess. The quality score, albeit being discrete, can be thought of as taking continuous values and the rounded value is given in the dataset. So, we postulate the following Bayesian linear regression model

$$\begin{aligned} y_i|\ \mu_i, \sigma &\overset{\text{ind.}}{\sim} N(\mu_i, \sigma^2), \\ \mu_i &= \beta_0 + X\beta, \end{aligned} \tag{1}$$

where X is the data matrix and β the vector of regression coefficients. To decide which of the parameters are relevant we use a non-informative prior, which allows us to use the R implementation of the linear model. We use the AIC to decide which parameters to keep. This results in the following parameters:

1. volatile acidity
2. chlorides
3. free sulfur dioxide
4. total sulfur dioxide
5. pH
6. sulphates
7. alcohol

This leads to the following model

$$\begin{aligned} &y_i|\ \mu_i, \sigma \overset{\text{ind.}}{\sim} N(\mu_i, \sigma^2), \\ &\qquad \mu_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \beta_4 x_{4,i} + \beta_1 x_{4,i} + \beta_5 x_{5,i} + \beta_6 x_{6,i} + \beta_7 x_{7,i}, \\ &\qquad \beta_i \overset{\text{iid.}}{\sim} N(0, 10^6),\ i = 0, \dots, 7 \\ &\qquad \sigma^2 \sim IG(0.5, 0.5) \end{aligned} \tag{2}$$

The priors are chosen in a way to obtain little information as, we have no prior conception of what the values should be. The model was fitted in R. As the diagnostics and plots showed a strong autocorrelation for some variables, we chose to iterate the MC chain for 500,000 steps and keep every 50th value. after taking these measures the trace plots and diagnostics looked fine.

The model assumptions seems to be met as the Q-Q plot, see Figure 2, indicates that residuals are approximately normally distributed and show a standard deviation of 0.64, which indicates a good fit, particularly considering that the *true values* we compare the model to are rounded to the nearest integer.

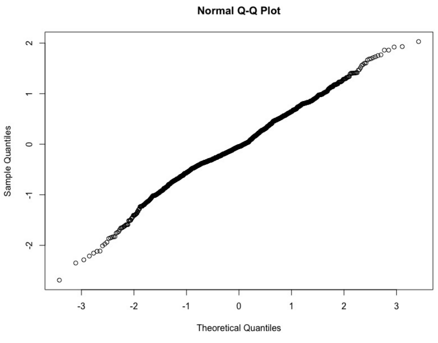

Figure 2: Q-Q plot of the residuals

We fitted a second linear model with only volatile acidity and alcohol, the two variables most strongly correlated with the quality, as explanatory variables. However, the DIC is higher than for the model described above, indicating a less adequate model.

4 Results

Looking at the posterior distributions, we see that all coefficient are significantly different from zero. We also see that the posterior probability that alcohol has a positive influence on the wine quality is basically 1. We want to predict the score of a wine with a feature vector of $(0.35, 0.073, 18, 38, 3.4, 0.87, 13.6)$. The mean posterior quality is 6.95 and the posterior probability that the wine is exceptional (score ≥ 7) is 17%.
However, one needs to be careful to use the model only for data similar to the given dataset as the mean posterior quality score for vodka (feature vector $(0, 0.01, 0, 0, 6, 0, 40)$) is an astonishing 13/10.

5 Conclusion

We learnt that in the 21st century humans do not need to bother themselves with deciding what wine they like when computers can do that more efficiently. Though it might come as a surprise to many when a vodka wins the best wine of the year award.

References

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis., J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553.