Predicting and classifying wines based on physical and chemical properties

January 8, 2019

Wine is a circa $300b industry, and somewhat unique in the modern age; whilst most consumer goods are specified and produced in a controlled manner with six sigma type methods, wine varies significantly not just between brands, but between batches.

The question set is:

"Chemically speaking, what types of wine are there? What predicts wine quality?"

This question comes in two parts: the latter is more traditionally suited to regression techniques, the former appears little more towards neural net approaches, although testing for collinearity and bimodals can still be of use.

The dataset consists of 6497 wines (1599 red, 4898 white) which have been assigned a quality rating (the dependent variable) from 0-10. 10 (independent) variables have been measured: fixed and volatile acidity, citric acid levels, residual sugar, chlorides, free and total sulfur dioxide, density, pH, sulphates and alcohol content (the dataset is complete, though units are not provided).

On a superficial level, there are clearly two types of wine: red and white, although that may be considered anthropomorphic. T-test of the two datasets confirms they are significantly ($\alpha < 0.01$) different in every measured variable, although alcohol content is marginal. Red wines have greater (fixed and volatile) acidity, chlorides and sulphates, and white wines have greater citric acid (132 red wines had no detectable citric acid at all), residual sugar, and (free and total) sulfur dioxide.

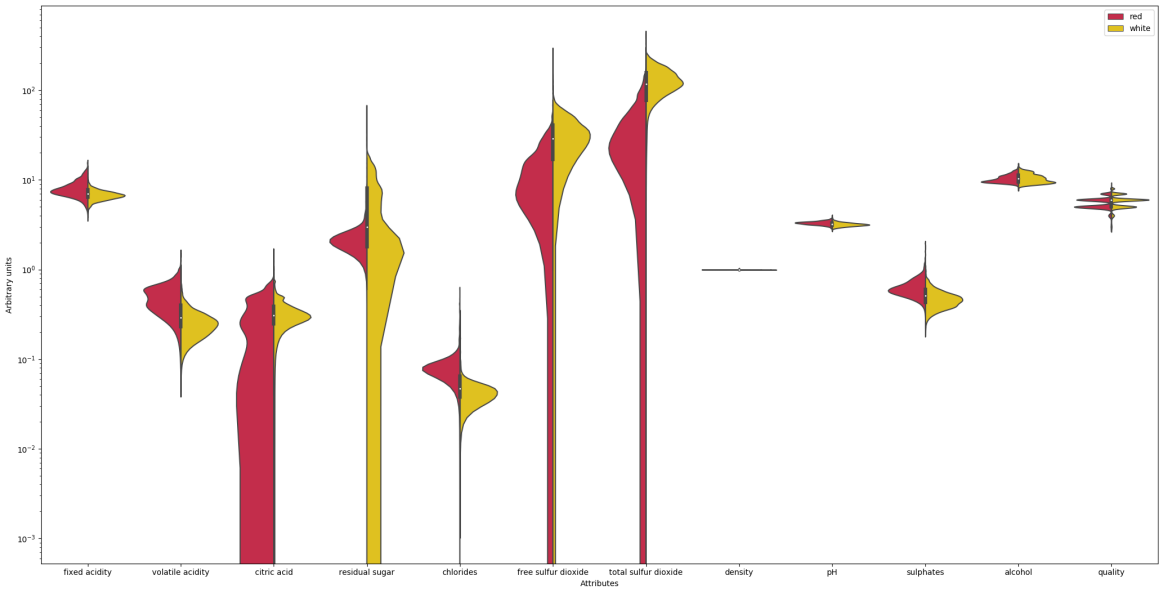

Figure 1: mean properties for red and white wines

Types of wine

llustrates correlation between different measured properties (darker colours indicate a greater association). Some correlations are intuitive; e.g. citric acid, fixed acidity and low pH are all correlated. Density is correlated with many factors in both red and white wines: this may also be intuitive as density is an aggregate property of the constituent chemicals. alcohol is correlated with a lower density (alcohol has a lower density than water), fixed acidity and residual sugars are correlated with higher density. For red wines only, chlorides are correlated with sulphates.

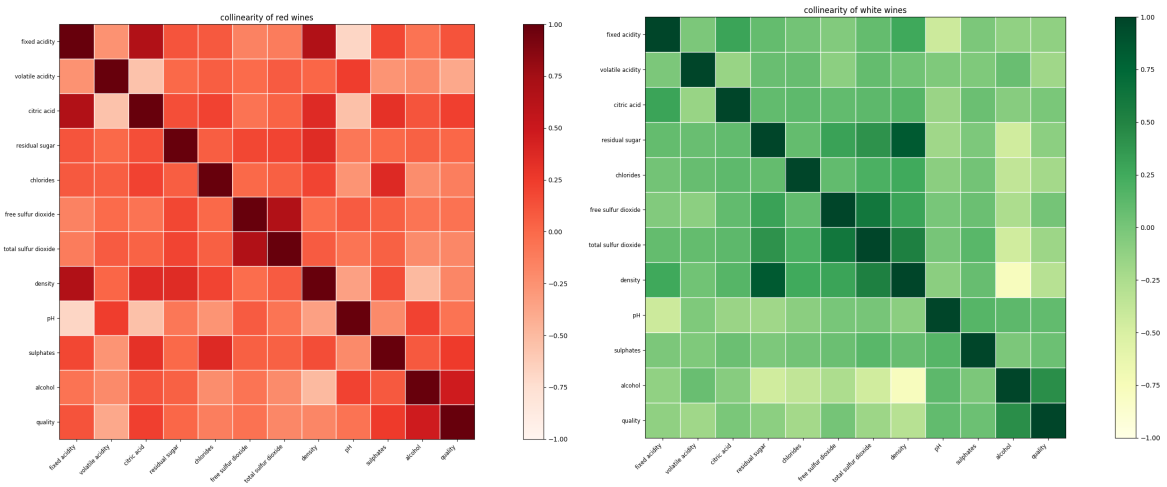

Figure 2: associations between different wine properties

Finding predictors

It would be reasonable to assume that that winemakers have roughly optimised each chemical present, so that the optimum lies somewhere within the space tested: in this case a quadratic fit is appropriate. It is equally possible that some quantities should simply be maximised.

Properties matching the stricter quadratic criteria are: for red wines, sulphates and citric acid; for white wines, free and total sulfur dioxide. For red wine only, increasing alcohol content increased perceived quality ($\alpha < 0.01$). For both types, volatile acidity, chloride and density should be minimised.

Quality red wines (scoring 6 or higher) have on average significantly ($\alpha < 0.01$) lower volatile acidity, higher citric acid (and therefore lower pH), higher sulphates, and higher alcohol, than poor reds (scoring 4 or lower). Quality white wines (scoring 6 or higher) have on average significantly ($\alpha < 0.01$) lower fixed and volatile acidity, higher residual sugar, lower chlorides, higher free sulfur dioxide but not total sulfur dioxide, lower density, than poor whites (scoring 4 or lower).
Both scored higher for higher alcohol content. Combining this with quadratic regressions, predictors of wine quality are summarised in figure 3.

	red	white
fixed acidity		LOW
volatile acidity	LOW	LOW
citric acid	HIGH	0.55
residual sugar		
chloride	LOW	LOW
free sulfur dioxide		0.0036
total sulfur dioxide	LOW	0.0021
density	LOW	LOW
pH		
sulphates	0.39	
alcohol	HIGH	

Figure 3: ideal properties for red and white wines