What determines good wine?

Project Proposal

Hien Le, Mahdiazhari Austian, Vera Grosskop

Machine Learning 2017 - Amsterdam University College
Supervisor: Gosia Migut

In this project, we will be using the dataset accessible from the UCI Machine Learning Repository (link: https://archive.ics.uci.edu/ml/datasets/wine+quality). This dataset was originally obtained from the paper “Modeling wine preferences by data mining from physicochemical properties” (2009) by Cortez et al.

The dataset consists of 11 features/attributes, all of which are physicochemical inputs as well as the label which is the score of the wine (score from 0 to 10). As these scores are whole integers, they can be seen as the 10 classes of the dataset. The used dataset contains 1599 red wine samples and 4898 white wine samples. We will compute the two wines separately, because the physicochemical consistency is different and a good prediction will not be possible if they are considered at the same time. Alternatively, an extra binary attribute could be introduced with the wine colour. The dataset contains more averagely scored samples than excellent or poor sample scores. The eleven attributes are as listed below:

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

The following table clarifies the range of the values in the data (Cortez et al., 2009).

Table 1
The physicochemical data statistics per wine type.

Attribute (units)	Red wine			White wine		
	Min	Max	Mean	Min	Max	Mean
Fixed acidity (g(tartaric acid)/dm^3)	4.6	15.9	8.3	3.8	14.2	6.9
Volatile acidity (g(acetic acid)/dm^3)	0.1	1.6	0.5	0.1	1.1	0.3
Citric acid (g/dm^3)	0.0	1.0	0.3	0.0	1.7	0.3
Residual sugar (g/dm^3)	0.9	15.5	2.5	0.6	65.8	6.4
Chlorides (g(sodium chloride)/dm^3)	0.01	0.61	0.08	0.01	0.35	0.05
Free sulfur dioxide (mg/dm^3)	1	72	14	2	289	35
Total sulfur dioxide (mg/dm^3)	6	289	46	9	440	138
Density (g/cm^3)	0.990	1.004	0.996	0.987	1.039	0.994
pH	2.7	4.0	3.3	2.7	3.8	3.1
Sulphates (g(potassium sulphate)/dm^3)	0.3	2.0	0.7	0.2	1.1	0.5
Alcohol (vol.%)	8.4	14.9	10.4	8.0	14.2	10.4

Classification problem. We want to classify a wine given its attributes into a quality (a class of score between 0-10). Moreover, after we have finished the classification step, if possible we also would like to create a boundary for good or bad wine (for example, a wine with a score 7 and above as good (1), and less than that as bad wine (0)).

We would also like to answer two questions:

- Which physicochemical properties in a wine are crucial for a good wine score?
- What are the right values for each of the features in order to get a good wine?

For this classification problem, we are going to start with logistic regression to predict the 10 classes. This is due to the fact that there are multiple features with continuous data. Furthermore, some features may also be weighted more heavily than others. This implies that we can also apply Support Vector Machine to get a large-margin classification, which potentially improves the performance. What can be done after this is to experiment with the Random Forest Algorithm: we set a cut-off value for each attribute as well as the score (e.g > 7 is good wine, between 4 and 7 is medium, and under 4 is bad wine). For both of the algorithms, we will split the dataset in three parts: 60% training set, 20% cross-validation set, and 20% testing set.

A comparison will be made by comparing the ROC Curves as well as looking at the AUC.
To tune the parameters (regularisation parameters, alpha, attributes to split on), we will draw the curve of errors against iterations.

References

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