UNIVERSITÀ DEGLI STUDI DI MILANO

DIPARTIMENTO DI INFORMATICA

SEDE DI CREMA

**Wine quality prediction**

A project report for Information Management course

Crema – 2015

# Introduction

An article “Modeling wine preferences by data mining from physicochemical properties” by Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos and José Reis published on sciencedirect.com in 2009 reviews and proposes a data mining approaches to predict wine taste quality evualations. The analysis is based on dataset which is a large compared to other taking in account the domain of the work.

The article reviews three techniques used for the predictions: the support vector machine, the multiple regression, neural network methods.

The support vector machine will be replicated in this work as an outperforming method in accuracy for this prediction. The naive Bayes classifier will be applied as an addition alernative classification method beside the SVM.

# Dataset

The dataset is decoupled to two separate CSV files. One of them contains samples of white wines and another of red wine. The dataset contains 12 numerous attributes:

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

The first 11 attributes are inputs which include physicochemical objective tests (e.g. PH values) and the 12th is output based on sensory data (median of at least 3 evaluations made by wine experts). Each expert graded the wine quality between 0 (very bad) and 10 (very excellent).

# Dataset investigation

To start investigate the data, some metrics were calculated for each attributes. The metrics involve mean, median, min, max, amplitude (max - min), standard deviation.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Mean | Median | Min | Max | Max - min | Standard deviation |
| **Fixed acidity** | 6.855 | 6.8 | 3.8 | 14.2 | 10.4 | 0.844 |
| **Volatile acidity** | 0.278 | 0.26 | 0.08 | 1.1 | 1.02 | 0.101 |
| **Citric acid** | 0.334 | 0.32 | 0 | 1.66 | 1.66 | 0.121 |
| **Residual sugar** | 6.391 | 5.2 | 0.6 | 65.8 | 65.2 | 5.072 |
| **Chlorides** | 0.046 | 0.043 | 0.009 | 0.346 | 0.337 | 0.022 |
| **Free sulfur dioxide** | 35.308 | 34 | 2 | 289 | 287 | 17.007 |
| **Total sulfur dioxide** | 138.361 | 134 | 9 | 440 | 431 | 42.498 |
| **Density** | 0.994 | 0.99374 | 0.98711 | 1.03898 | 0.05187 | 0.003 |
| **PH** | 3.188 | 3.18 | 2.72 | 3.82 | 1.1 | 0.151 |
| **Sulphates** | 0.49 | 0.47 | 0.22 | 1.08 | 0.86 | 0.114 |
| **Alcohol** | 10.514 | 10.4 | 8 | 14.2 | 6.2 | 1.231 |
| **Quality** | 5.878 | 6 | 3 | 9 | 6 | 0.886 |

Table 1: white wine metrics

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Mean | Median | Min | Max | Max - min | Standard deviation |
| **Fixed acidity** | 8.32 | 7.9 | 4.6 | 15.9 | 11.3 | 1.741 |
| **Volatile acidity** | 0.528 | 0.52 | 0.12 | 1.58 | 1.46 | 0.179 |
| **Citric acid** | 0.271 | 0.26 | 0 | 1 | 1 | 0.195 |
| **Residual sugar** | 2.539 | 2.2 | 0.9 | 15.5 | 14.6 | 1.41 |
| **Chlorides** | 0.087 | 0.079 | 0.012 | 0.611 | 0.599 | 0.047 |
| **Free sulfur dioxide** | 15.875 | 14 | 1 | 72 | 71 | 10.46 |
| **Total sulfur dioxide** | 46.468 | 38 | 6 | 289 | 283 | 32.895 |
| **Density** | 0.997 | 0.99675 | 0.99007 | 1.00369 | 0.01362 | 0.002 |
| **PH** | 3.311 | 3.31 | 2.74 | 4.01 | 1.27 | 0.154 |
| **Sulphates** | 0.658 | 0.62 | 0.33 | 2 | 1.67 | 0.17 |
| **Alcohol** | 10.423 | 10.2 | 8.4 | 14.9 | 6.5 | 1.066 |
| **Quality** | 5.636 | 6 | 3 | 8 | 5 | 0.808 |

Table 2: red wine metrics

The standard deviation quantifies the amount of variation or dispersion of a set of data values. A low standard deviation indicates that the data points tend to be very close to the mean (also called the expected value) of the set, while a high standard deviation indicates that the data points are spread out over a wider range of values. In our case white and red wine datasets have a high spreading for free sulfur dioxide attribute and even higher for total sulfur dioxide attribute. That

## Data correlation

In statistics, the correlation coefficient measures the strength and direction of a linear relationship between two variables (data attributes) on a scatterplot. The value is always between positive 1 and negative 1. Dataset attributes correlation coefficients and their interpretation:

* Exactly -1. A perfect downhill (negative) linear relationship
* -0.70. A strong downhill (negative) linear relationship
* -0.50. A moderate downhill (negative) relationship
* -0.30. A weak downhill (negative) linear relationship
* No linear relationship
* +0.30. A weak uphill (positive) linear relationship
* +0.50. A moderate uphill (positive) relationship
* +0.70. A strong uphill (positive) linear relationship
* Exactly +1. A perfect uphill (positive) linear relationship

There are no exactly perfect or even a strong correlation between data attributes, so we cannot to have accurate attributes calculations from other attributes after data reduction. This is one of the reasons why data reduction on attributes number was not performed.

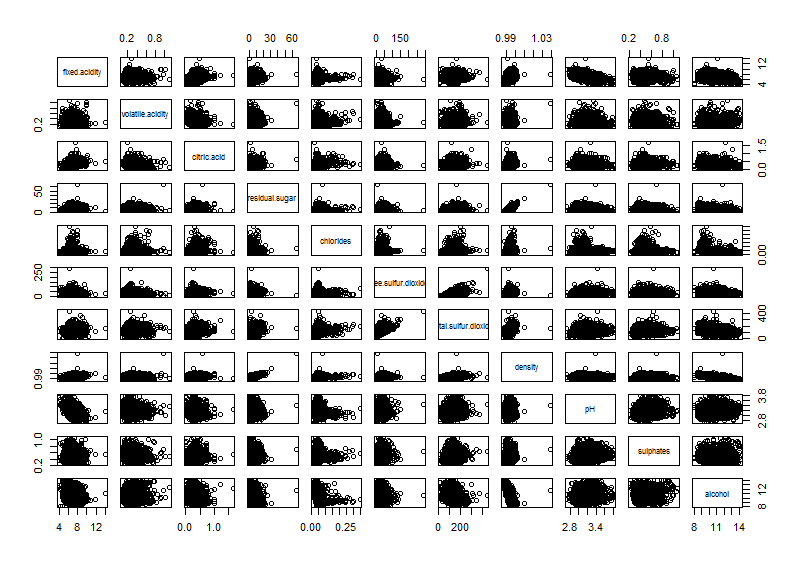


Figure 1: white wine correlation plots

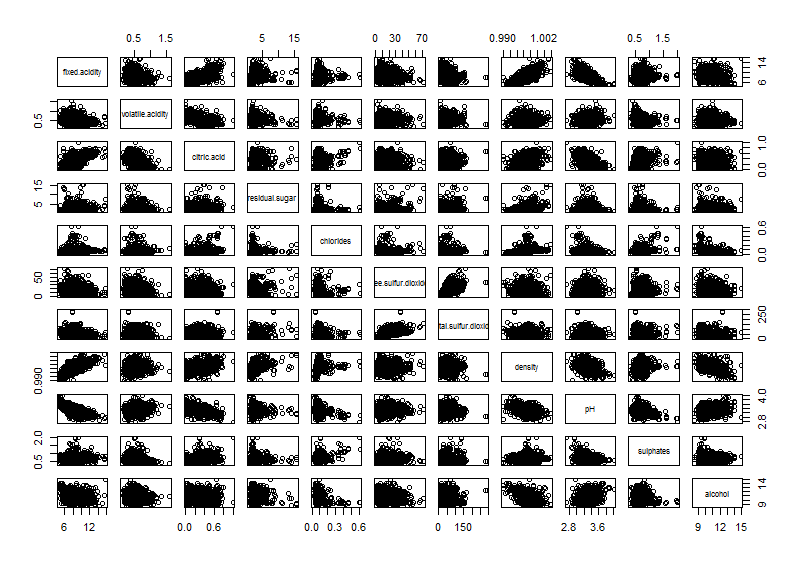


Figure 2: red wine correlation plots

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Fixed acidity | Volatile acidity | Citric acid | Residual sugar | Chlorides | Free sulfur dioxide | Total sulfur dioxide | Density | PH | Sulphates | Alcohol |
| **Fixed acidity** | 1.000 | -0.023 | 0.289 | 0.089 | 0.023 | -0.049 | 0.091 | 0.265 | -0.426 | -0.017 | -0.121 |
| **Volatile acidity** | -0.023 | 1.000 | -0.149 | 0.064 | 0.071 | -0.097 | 0.089 | 0.027 | -0.032 | -0.036 | 0.068 |
| **Citric acid** | 0.289 | -0.149 | 1.000 | 0.094 | 0.114 | 0.094 | 0.121 | 0.150 | -0.164 | 0.062 | -0.076 |
| **Residual sugar** | 0.089 | 0.064 | 0.094 | 1.000 | 0.089 | 0.299 | 0.401 | 0.839 | -0.194 | -0.027 | -0.451 |
| **Chlorides** | 0.023 | 0.071 | 0.114 | 0.089 | 1.000 | 0.101 | 0.199 | 0.257 | -0.090 | 0.017 | -0.360 |
| **Free sulfur dioxide** | -0.049 | -0.097 | 0.094 | 0.299 | 0.101 | 1.000 | 0.616 | 0.294 | -0.001 | 0.059 | -0.250 |
| **Total sulfur dioxide** | 0.091 | 0.089 | 0.121 | 0.401 | 0.199 | 0.616 | 1.000 | 0.530 | 0.002 | 0.135 | -0.449 |
| **Density** | 0.265 | 0.027 | 0.150 | 0.839 | 0.257 | 0.294 | 0.530 | 1.000 | -0.094 | 0.074 | -0.780 |
| **PH** | -0.426 | -0.032 | -0.164 | -0.194 | -0.090 | -0.001 | 0.002 | -0.094 | 1.000 | 0.156 | 0.121 |
| **Sulphates** | -0.017 | -0.036 | 0.062 | -0.027 | 0.017 | 0.059 | 0.135 | 0.074 | 0.156 | 1.000 | -0.017 |
| **Alcohol** | -0.121 | 0.068 | -0.076 | -0.451 | -0.360 | -0.250 | -0.449 | -0.780 | 0.121 | -0.017 | 1.000 |

Table 3: white wine correlation coefficients

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Fixed acidity | Volatile acidity | Citric acid | Residual sugar | Chlorides | Free sulfur dioxide | Total sulfur dioxide | Density | PH | Sulphates | Alcohol |
| **Fixed acidity** | 1.000 | -0.256 | 0.672 | 0.115 | 0.094 | -0.154 | -0.113 | 0.668 | -0.683 | 0.183 | -0.062 |
| **Volatile acidity** | -0.256 | 1.000 | -0.552 | 0.002 | 0.061 | -0.011 | 0.076 | 0.022 | 0.235 | -0.261 | -0.202 |
| **Citric acid** | 0.672 | -0.552 | 1.000 | 0.144 | 0.204 | -0.061 | 0.036 | 0.365 | -0.542 | 0.313 | 0.110 |
| **Residual sugar** | 0.115 | 0.002 | 0.144 | 1.000 | 0.056 | 0.187 | 0.203 | 0.355 | -0.086 | 0.006 | 0.042 |
| **Chlorides** | 0.094 | 0.061 | 0.204 | 0.056 | 1.000 | 0.006 | 0.047 | 0.201 | -0.265 | 0.371 | -0.221 |
| **Free sulfur dioxide** | -0.154 | -0.011 | -0.061 | 0.187 | 0.006 | 1.000 | 0.668 | -0.022 | 0.070 | 0.052 | -0.069 |
| **Total sulfur dioxide** | -0.113 | 0.076 | 0.036 | 0.203 | 0.047 | 0.668 | 1.000 | 0.071 | -0.066 | 0.043 | -0.206 |
| **Density** | 0.668 | 0.022 | 0.365 | 0.355 | 0.201 | -0.022 | 0.071 | 1.000 | -0.342 | 0.149 | -0.496 |
| **PH** | -0.683 | 0.235 | -0.542 | -0.086 | -0.265 | -0.001 | 0.002 | -0.094 | 1.000 | 0.156 | 0.121 |
| **Sulphates** | 0.183 | -0.261 | 0.313 | 0.006 | 0.371 | 0.059 | 0.135 | 0.074 | 0.156 | 1.000 | -0.017 |
| **Alcohol** | -0.062 | -0.202 | 0.110 | 0.042 | -0.221 | -0.250 | -0.449 | -0.780 | 0.121 | -0.017 | 1.000 |

Table 4: red wine correlation coefficients

# Data sets processing

The data sets of white and red wines have no missing values. Both datasets have some duplicated samples. These records has no influence in prediction of accuracy, but only slowing the performance of calculations, so it was removed leaving only unique data samples.

|  |  |  |
| --- | --- | --- |
|  | White wine | Red wine |
| **Total samples** | 4898 | 1599 |
| **Samples which have duplicats** | 5 | 5 |
| **Extra samples to be removed** | 937 | 240 |
| **Removed samples** | 19.13 % | 15.01 % |
| **Total unique samples left** | 3961 | 1359 |

Table 5: duplicated data statistics

# Quality prediction

For quality prediction was applied two different methods – SVM and naive Bayes classifier. Both methods implemented in R language with default configuration taking into account, that usually users using predefined configuration and this work aim is not to optimize these prediction models, but to implement them in practise and achieve satisfying results. The only exception was SVM method – a parameter for class classification was used in order to get quality ranks as classes, but not values to be calculated with a precision for decimals.

To evaluate the SVM and naive Bayes classifier, a robust 5-fold cross-validation was adopted. The data was divided into 5 partitions of equal size and one subset was tested each time while the remaining data are used for fitting the model. The process is repeated sequentially until all subsets have been tested.

## Confusion Matrix for red and white wines by SVM



Figure 4: white wine confusion matrix

Figure 3: red wine confusion matrix

## Confusion Matrix for red and white wines by Bayes



Figure 6: white wine confusion matrix

Figure 5: red wine confusion matrix

## Accuracy evaluation

* Mean absolute error (MAE) between actual and predicted data in the same units of them.
* Strict accuracy is calculated when the exact grade is predicted correctly.
* One-neighbor accuracy is calculated taking into account one grade error. If the true evaluation of quality is 5, 4 and 6 are still acceptable as a correct prediction.

## Results

The results shown that a high accuracy could be reached using SVM when one grade error is accepted as correct grade. This accuracy takes approximate 95-98%. Meanwhile naive Bayes classifier is outperformed by SVM, but still has a quite accurate prediction rate.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Red wine | White wine |
| SVM | Strict accuracy | 0.588 | 0.541 |
| One-neighbor accuracy | 0.949 | 0.979 |
| MAE | 0.471 | 0.434 |
| Bayes | Strict accuracy | 0.517 | 0.436 |
| One-neighbor accuracy | 0.915 | 0.915 |
| MAE | 0.529 | 0.702 |

## Conclusion

SVM and Bayes classifier were adopted in this implementation trying to check if data mining and these approaches can be applied in quite practical way. These predictions are nearly accurate to the quality evaluations rated by wine tasting specialists, so the prediction models could be used in order to help the specialists to specify the quality of wines.