Classification of wine quality from physicochemical properties

Konstantin Vinogradov

Abstract

This project explores different machine learning techniques while trying to predict human wine taste preferences based on physicochemical properties of wine. A large dataset of Portugal wine “vinho verde” is considered, it was splitted into two subsets for red and white wine samples. The task was approached as a multiclass classification problem and several standard classifiers were tested and tuned using grid search with cross-validation. The Support Vector Machine and Random Forest classifiers achieved equally good results for both subsets. Based on these findings a combining simple vote classifier was proposed and tested with 90% better performance on both datasets.

Motivation

Once viewed as a luxury good, nowadays wine is increasingly enjoyed by a wider range of consumers. Wine certification and quality assessment are key elements for development of the wine industry, they prevents the illegal adulteration and assures the wine quality. Wine certification is often assessed by physicochemical and sensory tests.

The development of an accurate model that could predict a sensory quality based on analytical data can be of great utility for the wine industry. On the one hand, such model can be very useful in the certification phase, since currently the sensory analysis is performed by human testers, being clearly a subjective approach. On the other hand, such a prediction system can also be useful for training oenology students and marketing purposes.

Dataset(s)

The data is collected from the UCI's Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Wine+Quality

The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. It has 1599 samples for red wine and 4898 for white wine.

Each sample contains 11 input variables based on most common physicochemical tests. The taste quality of the sample was evaluated by a minimum of three sensory assessors, by means of blind tastes, which graded the wine in a scale that ranges from 0 to 10, that matches to very bad to excellent quality, respectively. The final score is given by the median of these evaluations, which corresponds to the output variable.

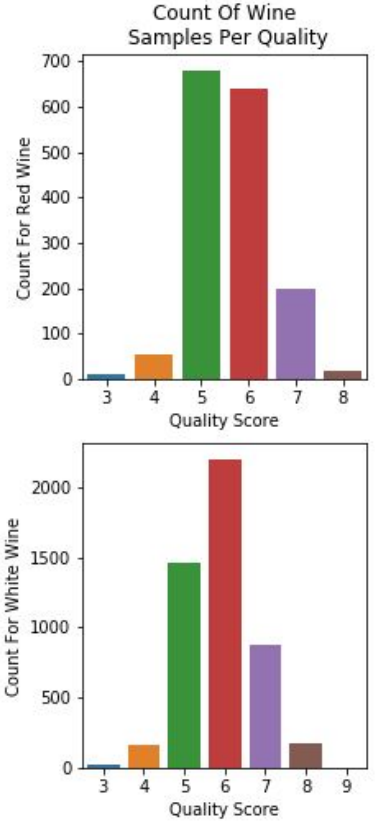
Data Preparation and Cleaning

Overall the Datasets were clean and ready for work.

However both Datasets are unbalanced - there are much more samples of average wines then excellent or very bad ones. As long as I was going to approach to the task as a classification one it could become a problem. Thus I've decided to merge the tails into one class:

For red: (3,4) => 4, (7,8) => 7

For white: (3,4) => 4, (8,9) => 8



Research Question(s)

1. Is it possible to predict wine taste preferences (taste quality) based on results of physicochemical tests?
2. What machine learning technique works best on these Datasets?
3. What impact on results is there from data standardisation and feature selection?

Methods

I adopt a classification approach to find a good classifier for prediction of taste preferences class based on input variables.

First of all I splitted Datasets into training (70% of samples) and test (30%) subsets. Then on training subsets I trained standard Classifiers: Support Vector Machine(linear and rbf), Logistic Regression, k-Fold Nearest Neighbor, Decision Tree and Random Forest. I tuned the global parameters using Grid-Search Cross Validation technique and did it on both Scaled and Unscaled datas.

Based on results I chose three classifiers with the best scores and proposed to use composite Median Voter Classifier (return median of prediction of SVM(rbf), Random Forest and 1-Nearest Neighbor), that showed better result comparing to any initial Classifiers.

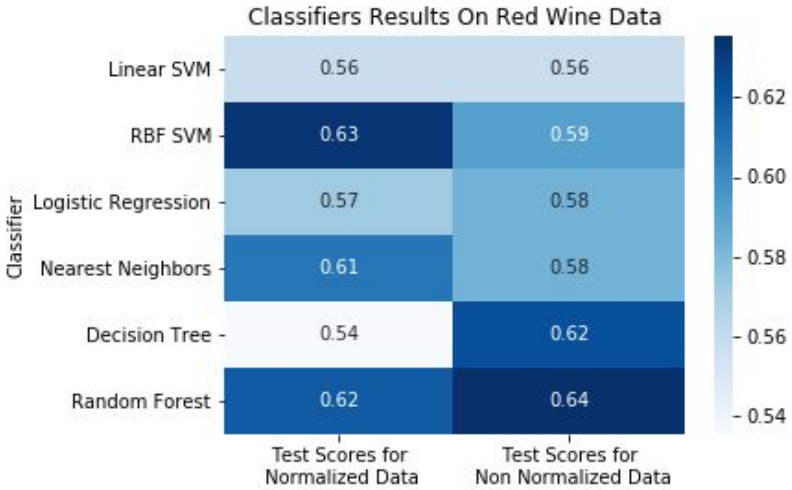
To be sure that results not happened due to randomness involved on many levels of the study I used multiple repetition technique. Based on it I compared distribution and main statistics of the results.

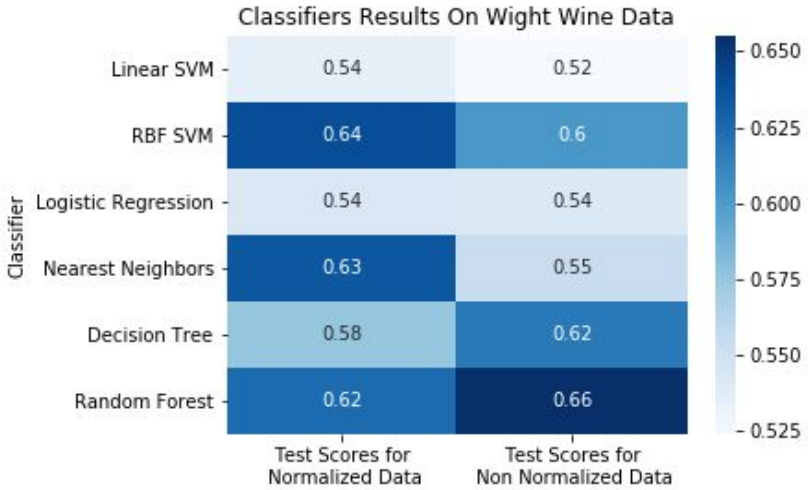
Findings

Firstly I checked how standard classifiers from Scikit-learn library could work on given Datasets. To get better results I used GridSearchCV class from the same library, which use cross validation technique to find best fitting global parameters for the classifiers. The results are introduced on the next slide and the main points are

- Several different classifiers achieved more then 60% accuracy
- As expected normalization of data is very important for Support Vector Machine and Nearest Neighbor and at least indifferent for Logistic Regression and Decision Tree based classifiers
- Best result on both Datasets get Random Forest classifier on non-scaled data followed by SVM-RBF on scaled data

Aggregated results of standard classifiers





Median Voter Classifier

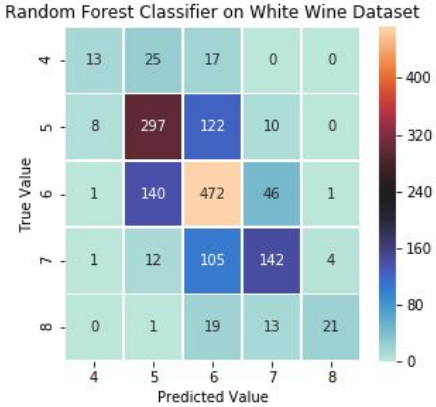
Based on the fact that there are several Classifiers getting results just over 60 percents it is logical to suggest existence of the way to improve the result by combining them. Thus I propose to use simple voting technique to boost accuracy of prediction.

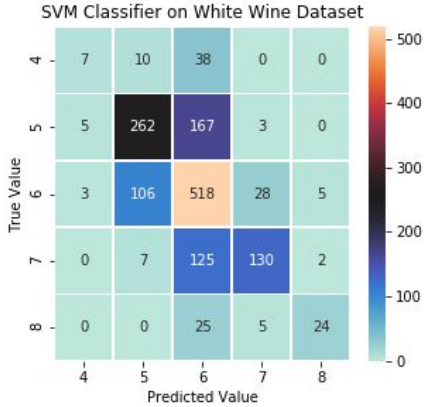
Median Voter Classifiers prediction is equal to median of Random Forest, SVM-RBF and Nearest Neighbor predictions. Median here looks like common sense choice - for three values median means that it takes either most voted value or less extreme prediction.

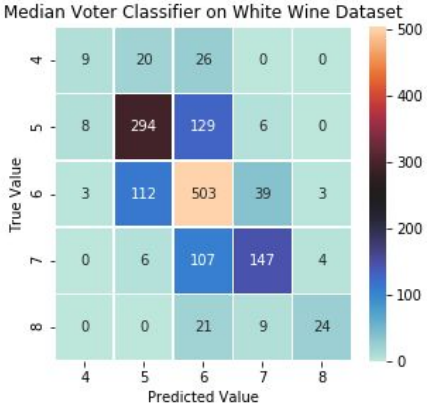
Thus defined Median Voter Classifier applied to given Datasets achieved better results: 66.46% on Red Wine Data and 64.17% on White Wine Data

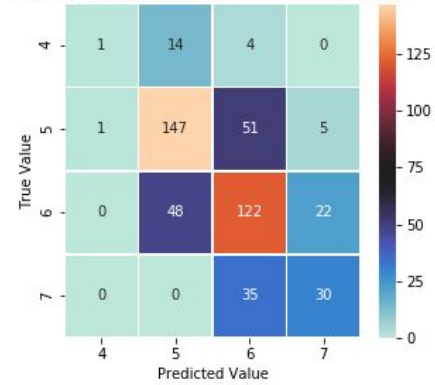
Let’s look at the confusion matrices of two best standard classifiers and new median voter classifier on the next slide.

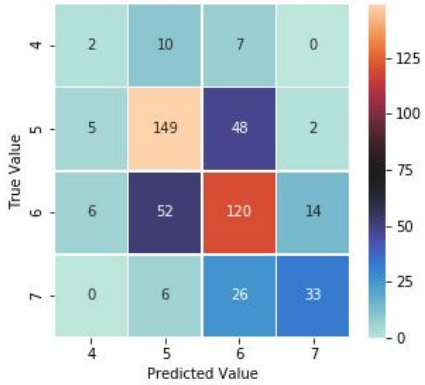
Confusion Matrices

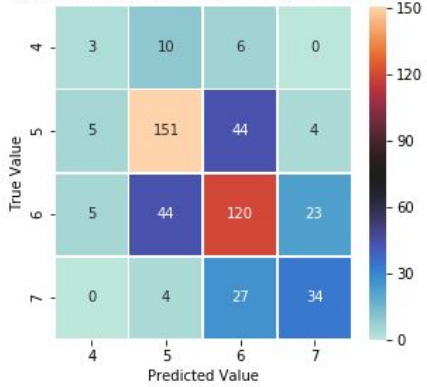












Confusion matrices observations

- Most of the values are close to the diagonals, denoting a good fit by the model based on any presented classifiers.
- The results of Median Voter Classifier are slightly better
- Based on confusion matrices it is possible to compute precision of the predictions for each predicted class. This statistic is important in practice, since in a real deployment setting the actual values are unknown and all predictions within a given column would be treated the same. In a table below introduced the results for Median Voter Classifier applied to White Wine Dataset. The results looks decent for exact predictions and excellent if prediction of the adjusted class is allowed.

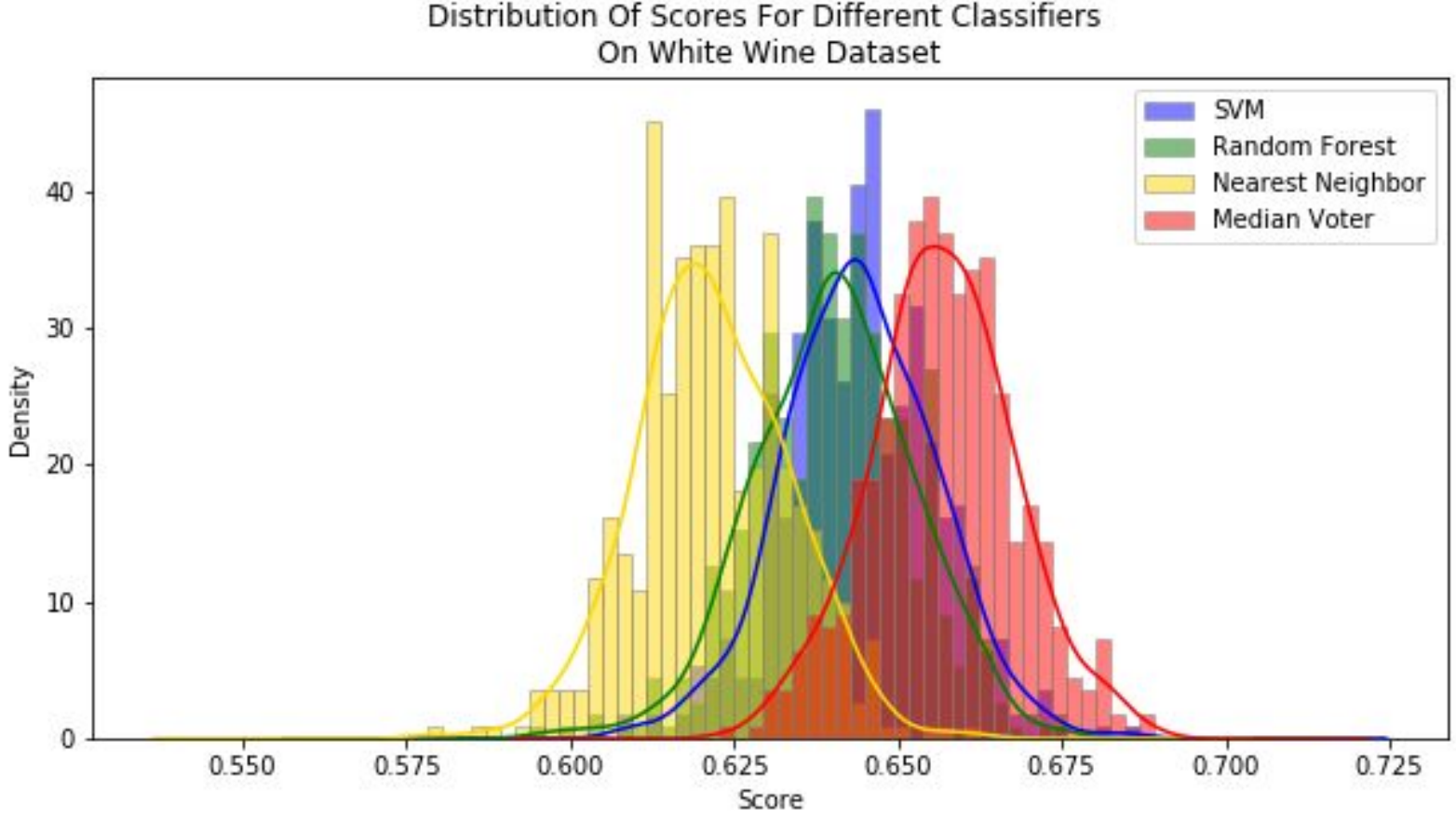
Predicted Quality Class:	**4**	**5**	**6**	**7**	**8**
Exact Predictions:	45.00%	68.06%	63.99%	73.13%	77.42%
Minimum Error Predictions:	85.00%	98.61%	94.02%	97.01%	90.32%

Statistical Evaluation of Classifiers

Considering the fact that the scores of SVM, Random Forest and Median Voter Classifiers are pretty close and there are a lot of randomness involved in the study no one can be sure which classifier is really the best one and how precision of prediction depends on random.

To address this problem I repeated computations described on previous slides 500 times for White Wine Dataset and 1000 times for Red Wine Dataset.

- Average scores of Median Voter Classifier were highest on both Datasets: 65.7% on White Data and 66.5% on Red Data
- Random Forest Classifier also shown very good result on Red Dataset and actually achieved the result better than Median Voter in 28% cases. Median Voter surpassed SVM on both Datasets and Random Forest on White Dataset in more than 90% cases.
- Distribution of scores achieved by different classifiers on White Wine Dataset introduced on the next slide


Distribution Of Scores For Different Classifiers
On White Wine Dataset
SVM
Random Forest
Nearest Neighbor
Median Voter
Density
0
10
20
30
40
0.550
0.575
0.600
0.625
0.650
0.675
0.700
0.725
Score

Limitations

The findings are only applicable to the Portuguese “Vinho Verde” wines it is possible that wines from different regions would have different characteristics. Moreover the given Datasets have quite unbalanced structure - there are much more decent wines and really small amount of excellent or very bad ones in the Datasets.

The Datasets themself have a bit questionable output variable. Given taste quality class isn’t actual quality class of the wine sample but just an estimation of it maden by several oenologists. To think of it they needed to take median of values given by different oenologists meaning even professionals tends to evaluate the same sample differently. The Datasets would consider two samples with values (4,5,5) and (5,5,6) as the same class 5, but in reality are they?

Conclusions

1. It is quite possible to predict wine's taste preferences based on physicochemical data. The introduced in this work Median Voter Classifier have shown around 66% accuracy of prediction for both Datasets. Moreover it gives at least 85% of precision for each predicted class given it errors no more than one class from the true one. Considering the remark about the Datasets I made in Limitations slide the result looks very promising.
2. The best results on these Datasets were achieved by Random Forest Classifier and Support Vector Machine. Nearest Neighbor technique also got a decent result, that allowed to use it in building a better custom classifier.
3. The results illustrated the importance of data standardization for such techniques as Support Vector Machine and Nearest Neighbor and indifference to it from Decision Tree based technique. Of course it's in line with the theory, but it always nice to see how it works in practice. Unfortunately I couldn't achieve any success in improving results by feature selection. That's why there is nothing about it in the findings.

Acknowledgements

The Datasets were taken from the UCI's Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Wine+Quality

Source of the data:

Paulo Cortez, University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez
A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal
@2009

References

1. 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. ISSN: 0167-9236.

   Available at: [@Elsevier] http://dx.doi.org/10.1016/j.dss.2009.05.016
   [Pre-press (pdf)] http://www3.dsi.uminho.pt/pcortez/winequality09.pdf
   [bib] http://www3.dsi.uminho.pt/pcortez/dss09.bib

2. Nebot, Àngela & Mugica, Francisco & Escobet, Antoni. (2015). Modeling Wine Preferences from Physicochemical Properties using Fuzzy Techniques. 501-507. 10.5220/0005551905010507.