Asia-pacific Journal of Convergent Research Interchange
Vol.3, No.4 December (2017), pp. 1-9
http://dx.doi.org/10.14257/apjcri.2017.12.07

Wine Quality Evaluation Using Machine Learning Algorithms

Xitiz Uniyal1), Prashant Barthwal2), Ashish Joshi3)

Abstract

There are many prediction systems available for problems like stock exchange, medical diagnosis, insurance calculation, etc. Wine Quality is one area where there is a big opportunity to recommend a good quality of wine to users based on their preferences as well as in historical data. This paper describes the work to learn and assess whether a given wine sample is of good quality or not. The use of machine learning techniques specifically the linear regression with stochastic gradient descent were explored, and the features that perform well on this classification were engineered. The main aim is to develop a cost-effective system to acquire knowledge using data analysis through machine learning algorithms to predict the quality of wine in a better way.

Keywords: wine quality, machine learning, supervised learning, linear regression, gradient descent.

1. Introduction

Globally, wine industry is nearly worth 300 billion dollars. Being able to predict the quality of wine would be very valuable addition to this industry. This paper aims to build a model that asses the wine quality for a given sample based on a given set of attributes. This is modeled by predicting the quality on a scale of 1 to 10 from a set of associated attributes. This is restricted to white wine segment as the required dataset is very good fit in this system but other wine type could also be used [1-5]. A way for this model to be used in practice is by including an automatic 'Predict: Quality' numerals message when a user enter a wine sample. The problem is modeled in a way that it is able to predict the wine quality numeral for any given sample.

In this endeavor, the Authors primarily delve into the following directions:

1) Implementation of linear regression by the use of stochastic gradient descent for generating predictions for the current data.
2) Generation of prediction for the multivariate linear regression.

Received(July 31, 2017), Review Result(1st: August 20, 2017, 2nd: October 13, 2017), Accepted(December 10, 2017)

[1](Corresponding Author) Dept. of Computer Science, THDC-Institute of Hydropower Engineering and Technology, Tehri, Uttarakhand, India
email: xitizuniyal@hotmail.com
[2]Dept. of Computer Science, THDC-Institute of Hydropower Engineering and Technology, Tehri, Uttarakhand, India
email: pbarthwal9@gmail.com
[3]Dept. of Computer Science, THDC-Institute of Hydropower Engineering and Technology, Tehri, Uttarakhand, India
email: a.joshi1986@gmail.com

There are systems that exist today that work with wine quality, but none of them model the problem in this way to asses the wine quality based on a given wine sample. To our knowledge, this is the first solution that attempts to predict a numeral in scale of 1 to 10 given a wine sample and attributes. One assumption in this work is that the attributes are not biased by the label, i.e., for a given wine sample all attributes are true and correct.

2. Contents

2.1 Data Collection

The data used in this project was obtained from UCI Machine Learning Repository[1]. The wine quality dataset is modeled using linear regression algorithm with stochastic gradient descent. The wine quality dataset contains the description of 4,898 white wines which includes the parameters as the acidity and the pH value [6-9].

The main aim is the prediction of the quality of wine on a scale ranging from 0 to 10. Dataset has to be normalized to the values ranging from 0 to 1 since every attribute has contrasting units for different scales. Baseline root mean squared error (RMSE) is achieved of 0.148 by using the Zero Rule Algorithm for prediction of the mean value on the dataset which is normalized. The way the problem is modeled is to learn from current data in order to make assessments.

2.2 Background Study

2.2.1 Multivariate Linear Regression

A straight line is used for modeling the relation between input and output values in Linear Regression [10-13]. The straight line can be considered as a hyper plane or plane of more than two dimensions. Combinations of various input values are used for predicting the output values. As per the following Eq. 1, the coefficient (b) is used to weigh the input attribute (x), and finding a set of coefficients that concludes in better predictions (y), is the aim of the machine learning algorithms.

$$y = b0 + b1 \ast x1 + b2 \ast x2 + ... \qquad) 1 ($$

2.2.2 Stochastic Gradient Descent -

The operation of reducing a function by gradients of cost function is known as Gradient Descent. In this process the form of cost as well as the derivative is well known so we can understand from any given point, which direction gradient can move. This can be the downhill towards the minimum value or uphill towards a maximum value. The stochastic gradient descent is used for minimizing the errors in the model of the data which estimates and modifies the coefficient after every iteration[2]. The working of this optimization algorithm is that every training instance is shown to the model one by one.

The calculation of error is done and the model is modified so as to minimize the error for the future prediction and this approach is repeated for the number of iterations. As per the following Eq. 2, the weight being optimized or the coefficient is the rate that we must configure. It is the predicted error of a model on a training data which is attributed to the weight and is the input value.

$$b = b - learning\ rate \ast error \ast x \qquad) \qquad 2$$

2.2.3 Making Predictions -

First, a function is made for the predictions as it will be needed for the projection of the coefficient values in the stochastic gradient descent. Moreover, it will also be required after the accomplishment of the model when predictions have to be done on the new set of the data. For a row of the coefficients, a predict function is used to predict the output value. The very first coefficient is known as the bias or b0 and is continually regarded as the intercept, and it is not at all liable for a definitive input value.

2.2.4 Estimating Coefficients -

The coefficient values of the training data can be decided using stochastic gradient descent. There are two criterions for stochastic gradient:

- Learning Rate: It is used to restrict the quantity of each coefficient being amended every time it is altered.
- Epochs: It is the count of working through the training data when the coefficients are amended.

Every coefficient for every row of the training data for an epoch is updated as per Eq. 2. The error of the model is the basis on which the coefficients are updated. Error is determined to be the difference in the prediction made and the output value which is expected.

$$error = prediction - expected$$

For every input attribute, one coefficient is to weigh, and the update is done in a regular manner.

$$b1(t + 1) = b1(t) - learning\_rate \ast error(t) \ast x1(t)$$

The coefficient which is present at the starting of the list also known as the intercept is modified in the same manner, but the input is left out since it has no relation with a peculiar input value.

$$b0(t+1) = b0(t) - learning\_rate * error(t)$$

If there is a continuous error drop in the final epoch, more epochs can be trained or the learning rate can be incremented in order that the quantity of the coefficients in each epoch is updated. In Python there is a predefined open source library, SciPy, so it cab be used for the operation on dataset. All the modules, objects and functions have to be imported. UCI Machine Learning repository is used for loading the data directly. The Pandas library is used for loading the data. Pandas library is also used to examine the data through data visualization and descriptive statistics.

2.2.5 Analysis of Dataset

During data loading, each column name is specified as it helps when there is a need to analyze the data. A comma separated value (CSV) file of the dataset can be downloaded and saved into the current working directory and it is loaded by the same method, by modifying the URL to the local file name[3]. The following steps are followed:

(a) **Load the libraries:** This step is used to import the libraries which are to be used.

(b) **Load the dataset:** The dataset which is to be evaluated is loaded with its filename.

(c) **Counting of the instances and attributes:** Instances are the rows and attributes are the columns which is approximately calculated by its shape property.

(d) **Taking a glance:** In this step the attribute's details is looked at which is consists of mean, minimum, maximum and percentile values.

(e) **Distribution of class:** The count of rows is taken into account.

After following these steps, the basic idea of the dataset that is to be used can be easily taken. Through visualization, the idea can be elongated for which two types of plots are used:

(1) **Univariate plots:** Used for understanding each and every attribute.

(2) **Multivariate plots:** Used for understanding the relationship among the attributes.

(f) **Creation of Whisker Plot:** To know the attribute distribution of the input values. [Fig. 2]

(g) **Creation of Histogram:** To get better understanding about distribution of each input variable and certainly it looks like the Gaussian distribution. [Fig. 3]

(h) **Creation of Scatter Plots:** Scatter plot matrix is basically created to see the input variables' structured relation. [Fig. 4]

2.3 Assessment

The algorithm is then applied on the real dataset. A linear regression model using stochastic gradient descent will be trained on wine quality dataset. The prime assumption is that in the current working directory there is a CSV copy of the dataset. At first the dataset is loaded, and then string values are converted to the numerical values. Every column then has to be normalized in the values ranging from 0 to 1.

Helper functions are used for loading and converting string values to float while preparing the dataset. Functions are used for the normalization of the dataset. K-fold cross-validation is used to evaluate the performance of the learned model on the data which is unseen that means that k-models will be constructed and estimated and evaluation of the performance will be done as mean model error. For the estimation of each model Root, the mean squared error is used. The cross validation split, RMSE metric, and evaluating helper functions provide the behaviors. Finally, the model is trained.

2.3.1 Steps for Wine Quality Prediction:

(a) Import the functional libraries: The libraries and functions required for assessment are imported.

(b) Load the file: The csv file is loaded.

(c) Conversion: The string column of the dataset is converted to the float type.

(d) Searching the values: The minimum and maximum value of each column is to be searched.

(e) Rescaling: The column dataset is rescaled in the range from 0 to 1.

(f) Splitting the dataset: The dataset is split into the k number of folds.

(g)Calculation of error: The root mean squared error is calculated.

(h) Algorithm Evaluation: The evaluation of algorithm is done through the cross validation split.

(i) Prediction: Prediction is made with the coefficients.

(j) Estimation of the coefficients of linear regression: The estimation is done through stochastic gradient descent.

(k) Loading and preparation of the data: The dataset is loaded.

(l) Normalization: The scores and the mean RMSE values are normalized.

2.4 Future Work

It is acknowledged that the problem being solved is demanding, specifically because of the following reasons:

1) Unclear predictability of quality: Any supervised learning model desire to gain from the labels, given the presented features. The underlying assumption that is made here is that the features that we have access to are ample to predict the quality. However one can figure out that the quality can depend highly on the experience that the user has that is not captured anywhere in the features. This could cause the correlation between the features and the label to be lesser than what would be optimal for a supervised

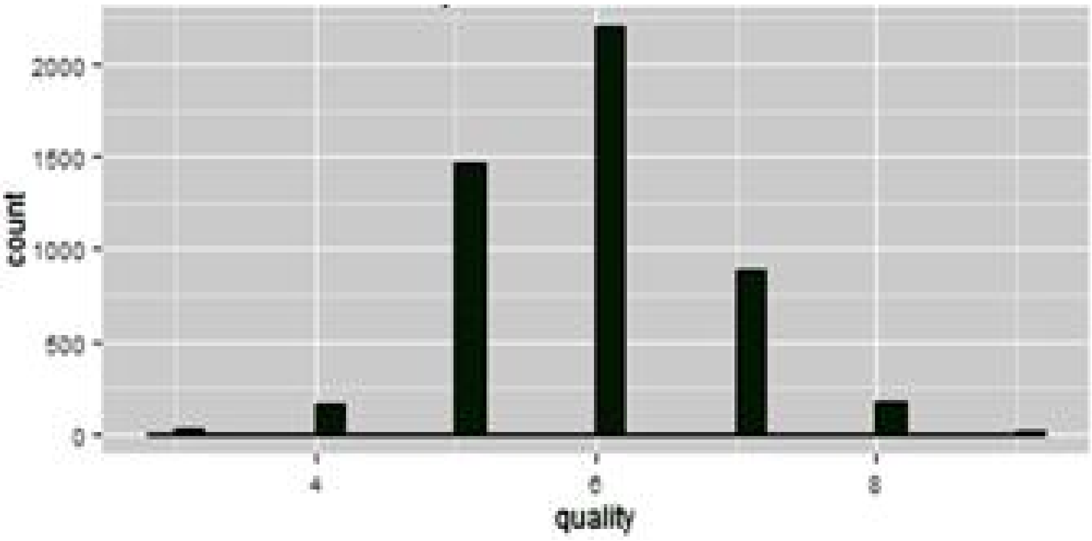
learning problem.

2) Bias in reviews used for training and evaluation: One hypothesis made is that a user's judgment to provide quality of a wine sample is random, and not biased by an unusually good or bad experience one has.

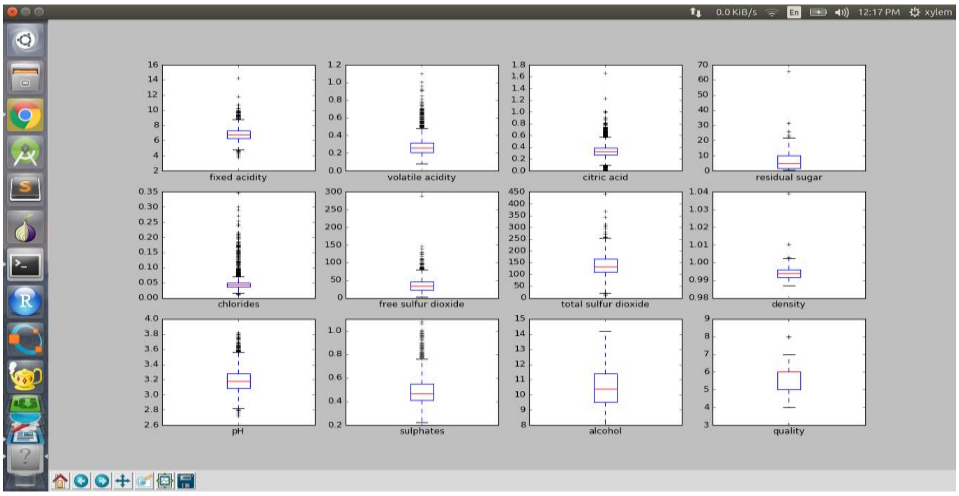
Future work would comprise trying to spot stronger features besides what is accessible in the datasets, as well as investing in an approach to collect training and evaluation data from alternate channel (such as explicit human judgment systems).

2.5 Tables and Figures

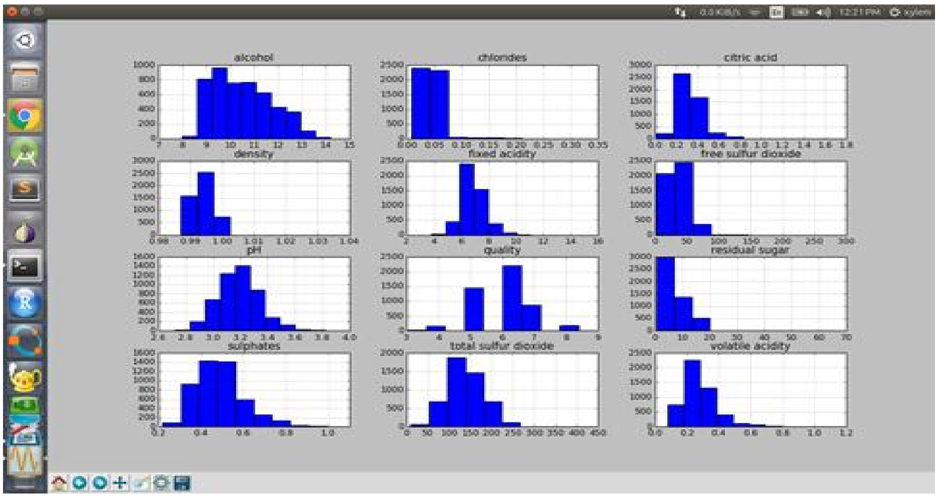
2.5.1 Figures



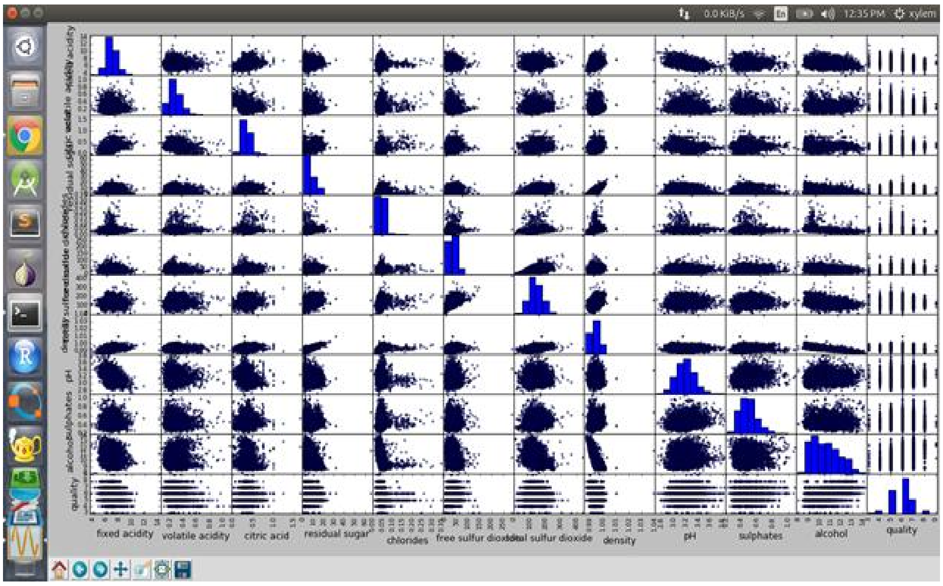
[Fig. 1] Wine Quality Ranking



[Fig. 2] Box and Whisker Plot



[Fig. 3] Histogram



[Fig. 4] Scatter Plot Matrix

2.5.2 Tables

[Table 1] The wine quality dataset contains the description of 4,898 white wines which includes parameters as:

Fixed Acidity	Volatile Acidity
Citric Acid	Residual Sugar
Chlorides	Free Sulfur Dioxide
Total Sulfur Dioxide	Density
pH	Sulfates
Alcohol	Quality

[Table 2] Fivefold division of data produced five score which our model generated based on the actual quality of wine and he predicted quality:

FOLD NUMBER	SCORE
1	0.12259834231519767
2	0.12733924130891316
3	0.12610773846663892
4	0.1289950071681572
5	0.1272180783291014

3. Conclusions

Result: Scores

[0.12259834231519767, 0.12733924130891316, 0.12610773846663892, 0.1289950071681572, 0.1272180783291014]

Mean RMSE: 0.126

Average RMSE of 0.126 was obtained from the proposed model which enables it to successfully assess and predict the quality of any given wine sample[4]. The number of folds for datasets could be varied to gain multiple insights about the results.

Acknowledgement

The National Digital Library of India was quite helpful in this project. In addition, we would like to thank Prof. Andrew Ng for the online lectures on insights when applying machine learning. We also want to thank Stanford University's CS229 Machine Learning online course and our peers for fruitful discussion on the

student forum.

References

[1] Wine quality dataset: https://archive.ics.uci.edu/ml/datasets/Wine+Quality

[2] T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning, 2nd ed., Springer, August (2008), pp. 9-101.

[3] S. Raschka, Python Machine Learning, Packt Publishing Limited, September, (2015), 454p.

[4] P. Wessa, Multiple Regression (v1.0.38) in Free Statistics Software (v1.1.23-r7), Office for Research Development and Education, (2015), http://www.wessa.net

[5] Ashish Joshi, R. H. Goudar, and Kanak Tewari, International Journal of Information and Education Technology (2012), Vol.2 (5): pp.480-482 ISSN: 2010-3689 DOI: 10.7763/IJIET.2012.V2.184

[6] Ashish Joshi, Kanak Tewari,Vivek Kumar and Dibyahash Bordoloi, International Journal Computer Science and Mobile Computing (2014), Vol.3, Issue.4, pp.1251-1258

[7] Ankit Kumar, Ashish Joshi, Anil Kumar, Ankush Mittal and D R Gangodkar, International Journal of Signal Processing, Image Processing and Pattern Recognition (2014), Vol.7, No.2, pp. 201-210

[8] Kanak Tewari, Ashish Joshi, Shalini Garg, RH Goudar, Dibyahash Bordoloi, 7th International Conference on Intelligent Systems and Control (ISCO) (2013), DOI: 10.1109/ISCO.2013.6481201

[9] RPubs - Wine Quality: Regression Models: https://rpubs.com/adena/wine

[10] Wine Quality Datasets: www3.dsi.uminho.pt/pcortez/wine/

[11] Predicting Wine Quality: web.stanford.edu/~ilker/doc/wine_Stats315A.pdf

[12] Ashish Joshi, Kanak Tewari, Ankit Kumar, Bhaskar Pant, 2013 International Conference on Information Systems and Computer Networks, DOI: 10.1109/ICISCON.2013.6524181,9-10 March (2013)

[13] Seunghan Lee, Juyoung Park, Kyungtae Kang, 2015 IEEE International Symposium on Systems Engineering (ISSE), DOI: 10.1109/SysEng.2015.7302752, pp.28-30 September, (2015)