Analysis and prediction of physicochemical factors of red wine on human taste

Yujia Ding
Computational Science, Mathematics, and Engineering
y5ding@ucsd.edu

Heng Zhang
Department of Electrical and Computer Engineering
hez009@ucsd.edu

ABSTRACT

The relationships between human sense and physicochemical factors of wine is an important task for the wine industry. However, there are a lot of physicochemical factors of wine and it is difficult to understand how those factors effect the human taste. We proposed to use data mining to solve this problem. Through feature analysis, we deleted four unrelated features. And then, the selected features were used to build three classification models based on three algorithms to predicted the 'good' wine for human taste. All three models achieved an accuracy rate of over 75%. According to our experimental results, SVM can performwell both on accuracy and BER, which has credible accuracy andavoids overfitting.

1 INTRODUCTION

Wine is a type of luxury in before, and it is affordable for a wider range of people nowadays. With the growth of wine market, there are two key points for the wine industry should be improved, which are wine production and wine selling. The wine certification is significant in both wine production and wine selling, because certification can prevent fake wine from entering the market and classify different wine for the pricing. The wine dataset used in this project is the red variants of the Portuguese *Vinho Verde* wine shown in Section 2.1 [2]. As the top ten exporting country of wine, Portugal's market share was 3.17% in 2005 and a type of its wine named *Vinho Verde* got 36% growth from 1997 to 2007 [2].

The process of general wine certification consists of two parts, which are physicochemical tests and sensory tests [3]. Physicochemical tests are based on physicochemical factors of wine such as alcohol, acidity, sugar, density and pH, while sensory tests are rely on experts' human sense like taste [8]. The human sense can not be quantified and specifically described, and the relationships between the human sense and physicochemical factors are complicated, so the wine classification is a complex task [7][8].

In order to understand the relationships between the sensory tests and physicochemical tests, data analysis technologies are able to used. The metadata contains a lot of useful information like trends, which is useful in improvement of decision making [11]. The method of finding useful high level information from metadata is called data mining, and there are several algorithms in the field of data mining like linear regression, support vector machines(SVMs) and logistic regression [6][9]. Linear regression is a type of linear approach to find the relationships between input features and output predictions [1]. In terms of probabilistic classifier of binary outputs such as good/bad, pass/fail and win/lose, the logistic regression model is suitable to make predictions [10]. If we want to achieve a non-probabilistic binary linear classifier, SVMs are useful algorithms [9]. When the data mining is applying, the selection of models and variables are significant for the whole process. The appropriate input variables will get accurate results, and a better performance will be achieved by discarding irrelevant variables [4]. The appropriate model can prevent overfitting of data, which means we can get more convincing results [5].

The main contributions of this work are as follow: (i) We show the distribution of different features, and effectively select features by analyzing their covariance, chi-square scores and the mean square errors obtaining from the linear regression model. It's effective to understand the influence of physicochemical factors of red wine on human taste. (ii) We conduct experiments to compare the performance of SVM model and Logistic Regression and KNN on predicting high quality red wine. It's helpful to understand the advantages and disadvantages of different models in classification prediction.

2 DATA ANALYSIS

2.1 Dataset

We use the red *vinho verde* wine data from UCI machine learning repository for our projects [2]. Dataset contains 11 different features of wine based on physicochemical tests, and the scores of wine quality from 0-10 based on sensory tests. The statistical description of 11 different physicochemical features is shown in Table 1. In terms of the scores of wine quality, the average of scores is 5.64 with 0.81 standard deviation, which means that most people give scores from 5-7. Therefore, this is not a balanced dataset, there are much more normal wines than good or bad wines.

2.2 Data Visualization

In order to understand all the distributions of features, we plot the histograms of corresponding distributions of features with their Cumulative Distribution Functions(CDF). Also, the scores of wine quality are analyzed by plotting histogram. We combined all figures into Figure 1.

2.3 Data Setting

When we apply the data mining to find the relationships between the human sense and physicochemical factors of wine, all different physicochemical features can be treated as input variables. Output predictions of model are the scores of wine quality from 0-10 based on sensory tests. In this project, the whole dataset is divided into two parts: training set and testing set. The training set accounts for 70% of the total, which is used to train the prediction model. The remaining 30% is used as the testing set to test the performance of model we achieved.

Table 1: The physicochemical features statistical description.

Attribute (units)	Min	Max	Mean	Median	Std. Deviation
Fixed acidity ($g(tartaricacid)/dm^3$)	4.60	15.90	8.32	7.90	1.74
Volatile acidity ($g(aceticacid)/dm^3$)	0.12	1.58	0.53	0.52	0.18
Citric acid (g/dm^3)	0	1	0.27	0.26	0.19
Residual sugar (g/dm^3)	0.90	15.50	2.54	2.20	1.41
Chlorides ($g(sodiumchloride)/dm^3$)	0.01	0.61	0.09	0.08	0.05
Free sulfur dioxide (mg/dm^3)	1	72	15.87	14	10.50
Total sulfur dioxide (mg/dm^3)	6	289	46.47	38	32.90
Density (g/cm^3)	0.99	1	0.99	0.99	0
pH	2.74	4.01	3.31	3.31	0.15
Sulphates ($g(potassiumsulphate)/dm^3$)	0.33	2	0.66	0.62	0.17
Alcohol ($vol.\%$)	8.40	14.90	10.42	10.20	1.07

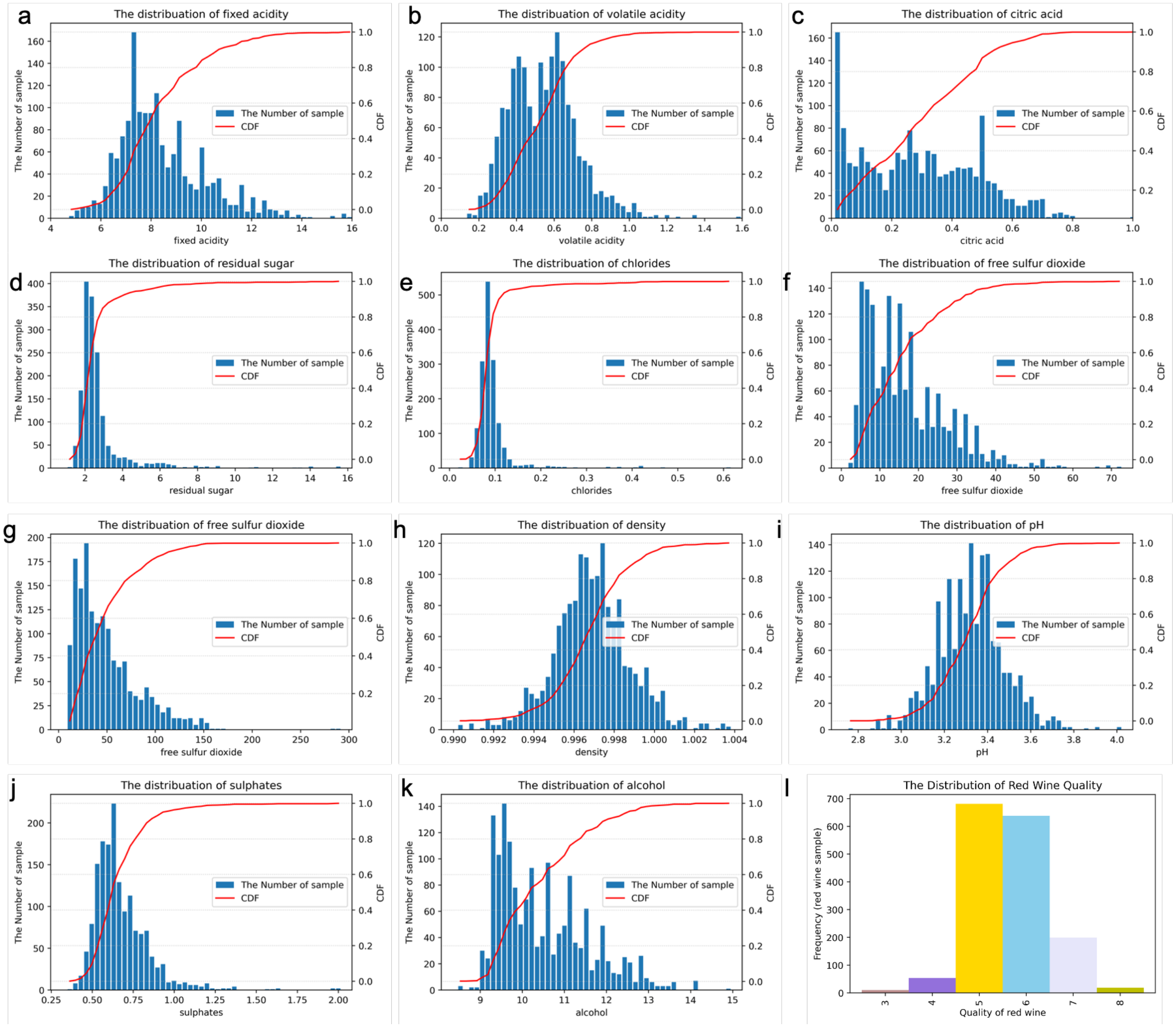

Figure 1: Data Visualization.

3 EXPERIMENTAL EVALUATION

3.1 Experimental Settings

3.1.1 Features Analysis. We use three methods to evaluate these features. First, we calculate the covariance between features and remove uncorrelated features. Meanwhile, we can get scores of features through chi-squared test. The trivial feature contributes less to prediction. We also apply the linear regression to figure out features that have an important impact on human sense, and then we select those important features for our classification task.

3.1.2 Prediction Task. In order to find the relationships between the human sense and physicochemical features, three classification models are built based on logistic regression, support vector machine(SVM) and K-Nearest Neighbor(KNN) by using features we selected before. In this project, the cutoff between 'good' and 'not good' is 6.5, so wine get 6.5 or higher score are classified as 'good'.

3.2 Data Mining Models

3.2.1 Linear Regression.

$$quality \cong \theta_0 + \theta_1 \times [feature_1] + ... + \theta_n \times [feature_n] \qquad (1)$$

3.2.2 SVM. For any real number, when $w^T x + b > 1$, $y_i = 1$, and when $w^T x + b < -1$, $y_i = -1$. Therefore:

$$min \frac{1}{2} \|w\|^2 \qquad (2)$$

$$s.t. y_i(w^T x_i + b) \geq 1, i = 1, 2, ..., m \qquad (3)$$

3.2.3 Logistic Regression.

$$p(good\ wine) \cong \sigma(\theta_0 + \theta_1 \times [feature_1] + ... + \theta_n \times [feature_n]) \quad (4)$$

3.2.4 k-Nearest Neighbor(KNN).

$$d(x, y) \cong \sum_{i=0}^{i=n} \sqrt{(xi - yi)^2} \qquad (5)$$

3.3 Evaluation Metrics

3.3.1 Features Analysis. We use Mean-squared error (MSE) to evaluate the performance of the each feature.

$$MSE = \frac{1}{N} \Sigma_{i=1}^{N} (y_i - X_i \cdot \theta)^2 \qquad (6)$$

3.3.2 Prediction Task. We use parameters from Table 2 to calculate Accuracy and Balanced Error Rate (BER) to evaluate the performance of the models.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \qquad (7)$$

$$BER = \frac{1}{2}(\frac{FP}{FP + TN} + \frac{FN}{FN + TP}) \qquad (8)$$

3.4 Implementation Detailsl

3.4.1 Features Analysis. We first normalize all features to be between zero and one to ensure the weight of each feature's influence is consistent.In calculating covariance between features, we collect all samples' features as a K*N matrix A (K is features of each sample and N is the number of samples) and subtract mean features of all samples. Then we calculate covariance of matrix A and set 'rowvar=True'. In chi-square test, we obtain the chi-square score of features. Meanwhile, we apply linear regression to find useful features. We first calculate the basic Mse by using all features, and then each feature is removed in turn to observe the change of MSE.

3.4.2 Prediction Task. We label both train and test data by the threshold 6.5. Train three classification models respectively. In logistic regression model, because this dataset is not balanced as mention in Section 2.1, thus we set 'class weight = balanced' to automatically adjust weights. In SVM model, except for set 'class weight = balanced', because the default kernel 'rbf' can easily get a overfitting result like an abnormally high accuracy, we change its kernel into 'poly'. In KNN model, we set 'n neighbors = 9' to get a better classification effect, and then set 'weights= distance' to adjust weights to eliminate effect of unbalanced dataset. Typically, the value of k can be the integer nearest to the square root of the number observation. However, for samples with high-dimensional features, the larger value of K may lead to overfiting.

3.5 Result

In the analysis of covariance, some features, such as volatile acidity and pH have negative covariance, indicating that they are negatively correlated with other factors. And density is an independent feature that has smallest covariance, only 2.6367e-05. And then the chi-squared test score come to similar conclusion: density is a trivial feature which obtains the lowest score (1.3111e-04). Applying linear regression to find suitable featureswe first obtain a small MSE (0.4498) by using all features to predict rating score, and then removing some features, such as density, total sulfur dioxide and citric acid, further reduce MSE as shown in Table 3. It shows that removing these features is beneficial to build linear regression model

Feature	MSE(subtracted)
Fixed acidity	5.9176e-05
Volatile acidity	0.03174
Citric acid	-5.4899e-05
Residual sugar	-0.001929
Chlorides	0.007706
Free sulfur dioxide	0.001597
Total sulfur dioxide	-0.0033965
Density	-0.001660
pH	0.001948
Sulphates	0.02323
Alcohol	0.02302

Table 3: Influence of features on MSE.

The performance in test set of three classification models are shown in Table 4. Note that the first two of these have similar performance in prediction task. SVM has better performance than logistic regression model, because SVM achieves a higher accuracy of 78.61% and the similar BER with logistic regression model. In terms of KNN model, it has the highest accuracy and the highest BER, which means this model has good performance of prediction of the negative samples. Thus, KNN model is easily influenced by unbiased dataset.

Table 2: Evaluating classifiers.

Prediction Label	1	0	Sum
1	True Positive (TP)	False Negative (FN)	Actual Positive(TP + FN)
0	False Positive (FP)	True Negative(TN)	Actual Negative(FP + TN)
Sum	Predicted Positive(TP + FP)	Predicted Negative(FN + TN)	TP+FP+FN+TN

Classification Model	Accuracy	BER
Logistic Regression	0.7654	0.2085
Support Vector Machine	0.7861	0.2106
K-Nearest Neighbor	0.8649	0.3377

Table 4: Prediction Results.

4 CONCLUSION

4.1 Discussion

In this project, we have selected several physicochemical factors of wine, which are related to scores of human sensory tests. Indeed, we proposed that those uncorrelated features can be deleted when we do prediction of human sense of wine, which can simplify the calculation of doing prediction. Also, we can get better represent the properties of the overall dataset by deleting those uncorrelated features. In the prediction task, all three classification models achieved an accuracy rate of over 75 percent, which meets our expectations because of feature selections. However, we do not suggest that those uncorrelated features should be deleted from the physicochemical test of wine, because those features are useful in checking for fake wine or harmful ingredients of wine.

4.2 Future Work

Based on this project, we can build a wine recommendation system by using more complicated algorithms. So far we have learned there are several features have significant influence on human sensory. Thus, the simple recommendation system can be achieved by recommending 'good' wine to users. Furthermore, if we can learned users' purchase records, we will understand the consumptive habits of users and will achieve a suitable recommendation system.

REFERENCES

[1] Jacob Cohen, Patricia Cohen, Stephen G West, and Leona S Aiken. 2013. *Applied multiple regression/correlation analysis for the behavioral sciences.* Routledge.

[2] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. 2009. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems* 47, 4 (2009), 547–553.

[3] S Ebeler. 1999. Flavor Chemistry—Thirty Years of Progress. *chapter Linking flavour chemistry to sensory analysis of wine* (1999), 409–422.

[4] Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *Journal of machine learning research* 3, Mar (2003), 1157–1182.

[5] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media.

[6] H Witten Ian and Frank Eibe. 2005. Data mining: Practical machine learning tools and techniques.

[7] A Legin, A Rudnitskaya, L Lvova, Yu Vlasov, C Di Natale, and A D'amico. 2003. Evaluation of Italian wine by the electronic tongue: recognition, quantitative analysis and correlation with human sensory perception. *Analytica Chimica Acta* 484, 1 (2003), 33–44.

[8] David V Smith and Robert F Margolskee. 2006. Making sense of taste. *SciAm* 16, 3 (2006), 84–92.

[9] Alex J Smola and Bernhard Schölkopf. 2004. A tutorial on support vector regression. *Statistics and computing* 14, 3 (2004), 199–222.

[10] Juliana Tolles and William J Meurer. 2016. Logistic regression: relating patient characteristics to outcomes. *Jama* 316, 5 (2016), 533–534.

[11] Efraim Turban, Ramesh Sharda, Jay E Aronson, and David King. 2008. *Business intelligence: A managerial approach.* Pearson Prentice Hall Corydonˆ eIndiana Indiana.