Wine Quality Prediction Through Linear Regression

Aguilar Valenzuela Sayde Alitzel

Abstract – A dataset of physicochemical factors on wine was analyzed to find its correlation with the quality of white wine. A linear regression model was proposed and obtained after a filtering of the variables.

I. INTRODUCTION

The preferences on wine are subjective decisions of wine pickers based on smell, taste, intensity, depth, complexity and other perceptual factors. However, an analysis of physicochemical factors will be done to determine how those affect the perception of the quality of the wine. (Vinetur 2007)

A dataset of wine quality taken from the UC Irvine Dataset Machine Learning repository will be analyzed to predict a model with a linear regression technique. The factors to be analyzed are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, density, pH, sulphates and alcohol.

A Pearson correlation analysis is previously performed to determine the impact of the variables on the quality of the wine, which can be useful for the wine producers to have knowledge on it. Also, the equation can be used to predict the perception of the quality of the wine when doing wine prototypes.

II. PEARSON CORRELATION

A PPMC (Pearson Product Moment Correlation) analysis was conducted on the white wine quality dataset. The linear correlation of each variable with the quality was obtained to determine which were the best variables to be used for the linear regression hypothesis. The PPMC is calculated with the following formulae were n is the sampling size, x and y are the implicated variables. (Glen n.d.)

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2)][n(\sum y^2) - (\sum y)^2)]}} \quad 1)$$

The PPMC was obtained with the Pandas module on Python. The correlations of the variables are depicted in the table, were the red variables were discarded.

	Quality
Fixed acidity	-0.113662
Volatile acidity	-0.1947
Citric acidity	-0.1136
Residual sugar	-0.0975
Chlorides	-0.2099
Free sulfur dioxide	0.008158
Density	-0.3071
pH	0.0994
Sulphates	0.0536
Alcohol	0.4355

III. LINEAR REGRESSION

A linear regression model was proposed. In the hypothesis function (2) the θs are the coefficients and the xs are the entering values for the physicochemical values of the wine.

$$\theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4 + \theta_5 x_5 + \theta_6 x_6 \quad 2)$$

x_1	Fixed acidity
x_2	Volatile acidity
x_3	Citric acidity
x_4	Chlorides
x_5	Density
x_6	Alcohol

IV. TRAINING MECHANISM

To train the model and get a small mean square error the Gradient Descent (formula 3) method was used to minimize the error for each coefficient θ_x of the hypothesis. In this method a coefficient α is used as a learning rate for every iteration of the learning process.

$$\theta_j = \theta_j - \frac{\alpha}{m}\sum_{i=1}^{m}[(h_\theta(xi) - y]\text{xi} \quad 3)$$

The training was made until a MSE of less than 0.65 was obtained considering no normalization made in the data.

Before the decrease of the variables with the Pearson Correlation method, all were used to determine the model, expecting an error of less than 0.8 (figure 1). After the selection of the six most correlated variables the same error was being

expected (0.8) and the behavior of the training process is shown in figure 2.

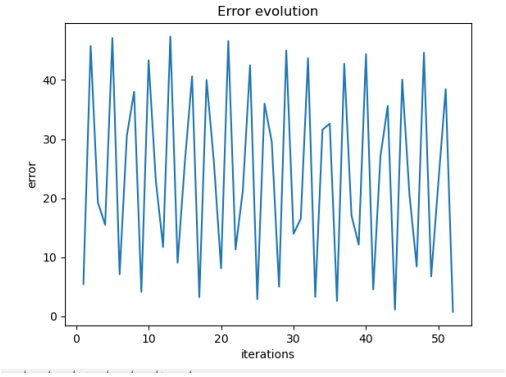

Figure 1 Error evolution for the model with 11 variables

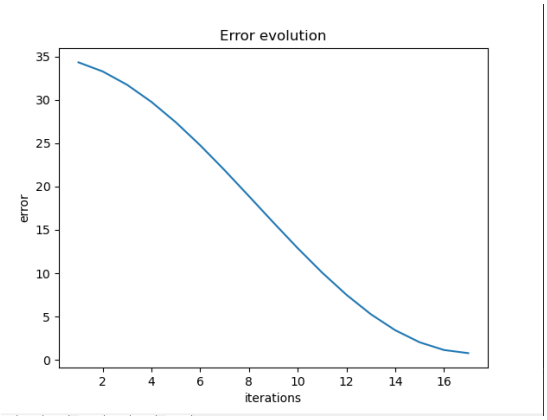

Figure 2 Error evolution for the model with 6 variables

From the figures can be observed the smaller oscillations to find the same error when discarding non correlated variables that made the model to be less convergent.

V. MODEL VALIDATION

The model was trained to get an error of less than 0.65. The obtained equation (4) was applied in the training instances to validate the model, obtaining an accuracy of 90.12, when using the equation to test with the non-training instances the accuracy increased to 90.25.

$$-0.1158 * x_1 - 1.4513 * x_2 + 0.534 * x_3 +$$

$$0.1241 * x_4 + 1.942 * x_5 + 0.4802 * x_6 \quad 4)$$

The accuracy was obtained with the division of the expected quality and the predicted quality, obtaining an average of the instances.

VI. CONCLUSIONS

When using linear regression to obtain a model is important to confirm if the variables are linearly related to the output, otherwise the model will not be accurate.

Diminishing the number of variables made possible to have a smoother error evolution and convergence of it.

VII. REFERENCES

Aguilar, Sayde. *GitHub.* 2020. https://github.com/aliaguilar/Wine-linear-regression.

Glen, Stephanie. "Statistics How to." n.d. https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/correlation-coefficient-formula/ (accessed March 22, 2020).

Paulo Cortez a, ∗ Ant´onio Cerdeira b Fernando Almeida b. «Modeling wine preferences by data mining.» *Universidad de Mino*, 2009.

«Vinetur .» 2007. https://www.vinetur.com/20180605 47361/5-formas-de-reconocer-un-buen-vino.html (último acceso: 22 de March de 2020).