Insights from Wine Quality: Linear Regression & Decision Tree Classifier

Kishan Shah, Prabhneet Arora and Shirish Pandagare

***Abstract*:** Wine Quality Data has been used to explore a linear regression fit to predict quality for white wine. Also, Binary Decision Tree was used to classify the two types of wines (red and white). The empirical error rate of the data distribution was computed to infer the true class probabilities (wine type) given the predictors of a specific wine. Furthermore, the application of linear regression model trained on white and tested on red was done as an attempt to draw insights into model reuse. Histograms depicting the difference in quality across the same models and model reused for a different dataset is used to draw concrete conclusions on model reuse.

Index terms: Linear Regression, Binary Decision Tree Classifier, Model reuse, Histograms

I. INTRODUCTION

Wine Quality data is a very popular dataset publicly used for research purposes with details being described by Cortez et al. [1] Vinho Verde is a unique product from northwest region of Portugal. Medium in alcohol, its specifically appreciated due to its freshness (especially in summer). [2] Two datasets pertaining to red and white "vinho verde" wine samples, from the north of Portugal have been used to test regression, using Linear Regression and Classification, using Binary Tree Classification methods.

A very interesting concept of model reuse has been introduced to create a generic and stable model for deployment in various circumstances in comparison to generating a wide variety of more specific and volatile models upon request. Based on this intuition, quality of the red wine was predicted on the linear regression model of white wine to obtain an RMSE value of XXX.

II. METHODOLOGY

A. Multivariate Linear Regression using Stochastic Gradient Descent on white wine dataset was used to optimize the estimated coefficients for the white wine to fit the white wine testing data with an accuracy of xxx. Stochastic Descent minimizes the cost function. At every step model makes a prediction, the error is computed, and the model is updated to reduce the error for the next prediction for a fixed number of iterations (epochs) using the formula as shown below:

$$\beta = \beta - learning\_rate * error * x$$

B. Binary Decision Tree Classifier using the combined dataset of red and white wine was used to classify the two wine types using the empirical error rate or the misclassification rate of the data distribution of XXX. The arguably low misclassification rate helped us to conclude that the two wine classes are significantly apart, and the decision tree model can correctly predict the wine quality XXX times.

C. White Wine Linear Regression with SGD tested on Red Wine: The white wine linear regression model was tested on red wine which resulted in RMSE (Root Mean Square Error) of XXX. RMSE is a frequently used criterion for measuring differences between estimated values (from our prediction model) and observed values. [3]

[Put the histograms' image]

III. RESULTS

As it is clearly visible from the histogram,

IV. CONCLUSION

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, 2009.

[2] C. (n.d.). Retrieved from http://www.vinhoverde.pt/en/

[3] Root-mean-square deviation. (2018, August 28). Retrieved from https://en.wikipedia.org/wiki/Root-mean-square_deviation