ECEN 689 - Applied Information Science
Challenge 4 - Wine Quality - Linear Regression and Decision Tree

Team 4 - Michael Bass, Mahalakshmi Sridharan, Ranjith Tamil Selvan

Introduction:

Challenge 4 involved understanding and applying the concepts of linear regression model and decision tree classifiers on the wine quality datasets of red and white wine. The challenge was divided into four parts.

Part 1: Linear regression model using Stochastic Gradient Descent was built using the white wine training dataset and it was utilized to predict the quality of the white wine test data.

Part 2: A decision tree classification algorithm was implemented to classify the combined wine quality dataset into red or white wine based on the 11 features.

Part 3: The linear regression model built using the white wine quality training data was used to predict the wine quality of the red wine test data.

Part 4: Based on the performance of the above, insights and conclusions were drawn.

Part 1: Linear Regression using SGD on White Wine Data:

A linear regression model using Stochastic Gradient Descent was built using the white wine quality training data and it was used to predict the wine quality of the white wine test data. The root mean square error using this model was around 0.76265 (from the Kaggle results). The following are the Beta coefficient vectors.

B0		0.03603904694
B1	fixed acidity	0.1146771955
B2	volatile acidity	-0.009396024859
B3	citric acid	0.008371766647
B4	residual sugar	0.0163466433
B5	chlorides	0.0006090529624
B6	free sulfur dioxide	0.01010376235
B7	total sulfur dioxide	-0.001031810263
B8	density	0.03558601666
B9	pH	0.129041084
B10	sulphates	0.02212728915
B11	alcohol	0.4119197471

Part 2: Performance of Decision Tree Classifier:

The classification of red and white wine was first implemented using decision tree classifier. Later, in order to improve the performance, it was replaced by random forest classifier which reduces overfitting and selects the most contributing features for classification. Random forest classification gave an accuracy of 0.99761 which was better than the decision tree classification accuracy of 0.99523. A highly precise classification was thus achieved.

Parts 3 & 4: Prediction Model Reuse and Performance:

The linear regression model trained using the white wine data was used to predict the quality of the red wine test set. This prediction was compared with the prediction accuracy/error that would have occurred if the red wine trained data was used to predict the red wine test data quality.

RMSE:

- (i) White wine trained model used to predict red wine quality - 0.7539651834500517
- (ii) Red wine trained model used to predict red wine quality - 0.7055570148576307

Based on the above RMSE values, we could see that the white trained model gave a RMSE score that was almost as good as the red wine trained model for the red wine testing set. The red wine linear regression model performed better than the white wine linear regression model which is as expected. However, the margin of better performance is very narrow. Considering that the decision tree classifier worked very well, we would expect that the red and white wines have drastically different features leading to a poor performance in the case of model reuse while prediction. Comparing the beta coefficients of the two trained models, we find that most of them have almost similar values except for two coefficients. This leads us to the inference that the classifier was able to perform well with just one or two vastly distinguishing features while the model reuse was able to perform well as the majority of the features had similar coefficient values.

		White wine trained model	Red wine trained model
B0		0.03603904694	0.0321854994
B1	fixed acidity	0.1146771955	0.1576736941
B2	volatile acidity	-0.009396024859	0.004990132
B3	citric acid	0.008371766647	0.0061189725
B4	residual sugar	0.0163466433	0.0187660046
B5	chlorides	0.0006090529624	0.0019177862
B6	free sulfur dioxide	0.01010376235	0.0126610122
B7	total sulfur dioxide	-0.001031810263	-0.0031607873
B8	density	0.03558601666	0.0319358392
B9	pH	0.129041084	0.1106724255
B10	sulphates	0.02212728915	0.0299558814
B11	alcohol	0.4119197471	0.3617163709

Conclusion:

The linear regression model using Stochastic Gradient Descent was trained on white wine and used to predict the wine quality of white wine and the model was reused for red wine quality prediction. The model reuse performanced well even though the classifier was accurate as the classification relies on few primary distinguishing features whereas the model utilizes all the features. Hence, a few significantly varying and the rest almost similar feature coefficients resulted in good performance of both the classifier and the model reuse.