Machine Learning I & Machine Learning II Project
Digital Text Analysis - University of Antwerp

Shima Rahimi 20214939

Dataset Description

The dataset is related to red variants of the Portuguese "Vinho Verde" wine [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.). The classes are ordered and not balanced (e.g. there are much more normal wines than excellent or poor ones).

First, we tried to understand the dataset, what each column signified and how it contributed in determining the quality of wine. Input variables (based on physicochemical tests) are as follows**:** 1 - fixed acidity, 2 - volatile acidity, 3 - citric acid, 4 - residual sugar, 5 – chlorides, 6 - free sulfur dioxide, 7 - total sulfur dioxide, 8 – density, 9 – pH, 10 – sulphates, and 11 – alcohol (each variable’s description has been explained within the code notebook).

The only output variable is column 12 of the wine dataset which is the quality (score between 0 and 10) – a discrete and categorical variable in nature. Quality score scale ranges from 1 to 10: 1 being the poorest and 10 being the best. Quality ratings 1, 2, 9 and 10 were not given by any observation, the only existing scores are from 3 to 8.

Then we visualized the data from statistical point of view. Tried to find anomalies and eliminate them. There were 240 duplicated values; but, because many different wines have the similar features we were not sure whether they were the same wine so we just ignore the duplicates. We checked the data for missing values and since there were no null values, we skipped the imputing functions during the application of the Classical Machine Learning methods.

Methods

Classical Machine Learning Models

We allocated 80% of the data for training and 20% remained for testing. Four different classification models were applied: Logistic Regression, Decision Tree Classifier, Random Forest Classifier, and Support Vector Classifier. Among the four classifiers, Random Forest Classifier yielded the highest accuracy score as shown in the following table:

	Model	Accuracy
1	Random Forest Classifier	0.69
2	Support Vector Classifier	0.60
3	Logistic Regression	0.59
4	Decision Tree Classifier	0.58

Based on these results, Random Forest Classifier was chosen for hyper parameter tuning. We tuned the model with GridSearchCv function and the final accuracy appeared to be 0.68.

Artificial Neural Network Models

For the neural models, we divided the data into 75% train set and 25% test set and development set (each set 12.5%). An early stopping regularization with a patience of 8 is applied to avoid overfitting. The neural model consists of an input layer, 3 dense layers with “ReLU” activation function, 1 flatten layer, and 1 dense layer with “SoftMax” activation function as the output layer. The loss function is “sparse_categorical_crossentropy”. The model got fitted with 50 epochs. The loss function and accuracy of the model has been plotted. A steady decrease in the loss function plot and a good amount of increase in the accuracy plot has been depicted.

Results

Model	**Accuracy**
Classical Model (Random Forest Classifier)	0.68
Artificial Neural Network Model (Sequential Keras)	0.51

As expected, the various properties in the dataset play a role in the quality of wine**.** Both classical machine learning and artificial neural network models were capable of predicting the quality with medium performance. We believe choosing the right parameters in hyperparameter tuning will increase the scoring of the classical model. The models have room for improvement with further optimizations, and gathering more data.

References

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.