Machine Learning Engineer Nanodegree

Red Wine Quality

Ng Hai Ming
April 21st, 2019

Domain Background

Out of various human senses, taste itself is the least understood [1], complicated by the fact it can be a highly subjective evaluation exercise. There are even instances where changes in stated wine prices influenced the volunteers' perception of its tastes as well as the humans' brain regions associated with experience of pleasure [2]. The ability to improve the objective identification of wine quality via supervised learning algorithms can facilitate the eventual stratification of various wine quality classes and the concomitant provision of objective & highly visible benchmarks to set fair prices of wines against [1].

Problem Statement

This project aims to use publicly available data from the UCI machine learning repository to classify the quality of red variants of Portuguese "Vinho Verde" wine. An arbitrary cut-off quality (sensory output variable) value of 7 would be used to determine if the wine is of good quality or not. The dataset would first be split into training and testing sets using sklearn's cross_validation train_test_split. During training, the performance of the model would be evaluated using fbeta_score of 0.5 and accuracy_score. Once trained on the training dataset, it would be used to predict the wine quality on the testing dataset. The objective of this exercise is to train a selected algorithm such as DecisionTreeClassifier to generalise well on the testing dataset, successfully classifying it with high accuracy and/or high F-score.

Datasets and Inputs

Data for the project, as attached with this proposal submission, can be downloaded from https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009. The dataset contains 1599 variants of the Portugese "Vinho Verde" wine, which includes 11 input physicochemical properties and 1 output sensory variable.

The input physicochemical properties includes [1, 3]:

1) Fixed Acidity: Tartaric Acid content, which do not evaporate quickly
2) Volatile Acidity: Acetic Acid content, which at too high levels can lead to unpleasant and vinegar taste
3) Citric Acid: Citric Acid content, which can add "freshness" and "flavours" to wine

4) Residual Sugar: The amount of sugar remaining post-fermentation, anything exceeding 45 g/L would be considered sweet
5) Chlorides: The amount of salt present in the wine
6) Free Sulfur Dioxide: Equilibrium concentration of free sulfur dioxide, which helps prevent microbial growth and oxidation of wine
7) Total Sulfur Dioxide: Amount of free and bound forms of sulfur dioxide, which remains undetectable in wine in low concentrations. Above 50ppm, it becomes evident in the smell and taste of wine
8) Density: The density should be close to that of water, varying from it depending on alcohol and/or sugar content
9) pH: Description on acidity of wine, which can influence the perception of tastes, since acidity leaves a sour taste and alkalinity a bitter taste
10) Sulphates: A wine additive that helps contribute to SO_2 levels, which has antioxidant and antimicrobial properties as discussed in *point 6*
11) Alcohol: The alcohol percent in the wine

The output sensory variable consists of [1, 3]:

1) Quality: The sensory analysis performed by a minimum of 3 sensory assessors (human tasters) using blind tastes, based on their expert experience and knowledge, on a scale ranging from 0 to 10

Solution Statement

This project will aim to classify the testing dataset accurately, which would consists 10% of the original dataset of 1599 variants of Portugese "Vinho Verde" wine in randomised fashion. It would predict if the wine is of good quality (quality >= 7) or not (quality < 7).

Benchmark Model

The performance of the model will be evaluated against the original paper, which utilises Neural Network and Support Vector Machines architecture, for overall accuracies figures for this dataset. There would be some degree of differences however, since the original paper retains the scale of 0-10 for the classification task, whereas in my proposal I've proposed a binary classification task with an arbitrary cut-off value of 7 (for the output quality figures) for the determination of a "good wine". The goal would be to match the highest accuracy figure of **89.0%** achieved in the Support Vector Machine used in the paper [1].

Evaluation Metrics

The following two evaluation metrics would be used:

1) F-beta score (score = 0.5) [4]:

$$F_\beta = (1 + \beta^2)\frac{precision\ .\ recall}{(\beta^2\ .\ precision) + recall}$$

$$F_{0.5} = (1 + 0.5^2)\frac{precision\ .\ recall}{(0.5^2\ .\ precision) + recall}$$

2) Accuracy Score [5]:

$$Accuracy = \frac{True\ Positives + True\ Negatives}{True\ Positives + True\ Negatives + False\ Positives + False\ Negatives}$$

Project Design

This project would be tackled in 4 different steps, namely:

1) Loading and Preprocessing of Data

   The data would be pre-processed with the conversion of output sensory (quality) figures into binary classification (0 for "not good" where quality < 7, 1 for "good" where quality >= 7) and the performance of min-max scaling on the 11 numerical input physicochemical properties.

2) Model Creation

   Three supervised learning algorithms (as of now, AdaBoostClassifier, DecisionTreeClassifier and Gaussian-NaïveBayes are proposed) would be employed to test on 1%, 10% and 100% of training data and then cross-validated on the entire testing set. The resulting bar plots generated would be plotted and the algorithm deemed to perform the best, in terms of the evaluation metrics used, would be elected as the algorithm of choice used for the final model creation.

3) Model Prediction (using Algorithm of Choice)

   The selected algorithm of choice would undergo the same procedure above, though with hyper-parameter tuning, to improve the resulting model's accuracy and f-score values. The hyper-parameter tuning would be either done through GridSearchCV or RandomSearchCV, dependent upon the hyper-parameters search-space as well as the selected algorithm's computational complexity.

4) Identification of important features

   To provide insights into the dataset, the sklearn's features importance library would be used to identify the important physicochemical properties, ranked in order of importance, that contribute to the quality of red wine. Also, Principal Component Analysis (PCA) could be used for the dimensionality reduction of the dataset and the selected algorithm of choice can also be ran on to investigate the effects such dimensionality reduction has on the original f-score and accuracy values.

Bibliography

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reise, "Modeling wine preferences by data mining from physicochemical properties," *Elsevier,* p. 7, 2009.

[2] K. Svitil, "CalTech," 14 January 2008. [Online]. Available: https://www.caltech.edu/about/news/wine-study-shows-price-influences-perception-1374.

[3] "Red Wine Quality," Kaggle, [Online]. Available: https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009. [Accessed 21 April 2019].

[4] "F1_Score," Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/F1_score. [Accessed 21 April 2019].

[5] A. Long, "Understanding Data Science Classification Metrics in Scikit-Learn in Python," 6 August 2018. [Online]. Available: https://towardsdatascience.com/understanding-data-science-classification-metrics-in-scikit-learn-in-python-3bc336865019. [Accessed 21 April 2019].