Predicting Type and Quality of Wine From Physical Properties

Nerses Alikhanyan

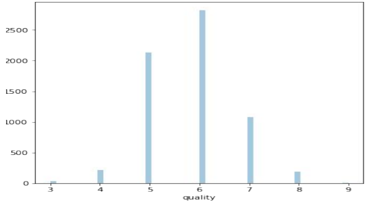
Applied Statistics and Data Science, Yerevan State University

Abstract

Classification and rating of wines is interesting and time consuming task performed by highly trained experts. To move beyond our reliance on human experts several automation techniques can be applied once objective quantitative data of wine features is identified. This paper describes various machine learning approaches to solve the task of wine type and quality classification. Note that data of wine features used is obtained from objective physical measurements only.
Code for experiments is available at https://github.com/mr-Hades/wine_classification

Dataset

The dataset considered here is vinho verde [1] which includes 6497 samples. It is unevenly split between two styles: 75 % of white and 25% of red wines. 11 physicochemical features describe each sample. Those are fixed acidity, volatile acidity, citric acid, concentration, residual sugar content, chlorides concentrations, free sulfur dioxide content, total sulfur dioxide content, density, pH, sulphates, concentrations, alcohol content.



GIven figure shows strong imbalance in quality classes. This is resolved by binning smaller classes together.

Methods

- Logistic Regression
- Support Vector Machine
- LightGBM [2] best model. Fast tree based gradient boosting library.
- TPE algorithm was used for hyper-parameter tuning

Results

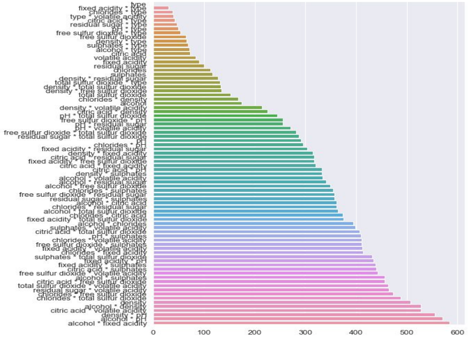
Metrics on Test set

Method	Add. Features	Balanced Acc.	F1
Logistic Regression	no	0.501	0.532
SVM (one-to-one)	no	0.592	0.595
LightGbm	no	0.64	0.649
Logistic Regression	yes	0.464	0.527
SVM (one-to-one)	yes	0.58	0.584
LightGbm	**yes**	**0.663**	**0.668**

Per Class F1 scores

Class	SVM	LightGBM	LightGBM (Add.)	Support
0	0.283	0.383	**0.432**	49
1	0.634	0.693	**0.712**	426
2	0.640	0.671	**0.687**	564
3	0.488	0.543	**0.585**	215
4	0.39	0.384	**0.508**	39

Feature Importances



References

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, “Modeling wine preferences by data mining from physicochemical properties,” Decision Support Systems, vol. 47, pp. 547–553, nov 2009.

[2] G. Ke, Q.Meng, T. Finely, T. Wang, W.Chen, W. Ma, Q. Ye, T. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree", Advances in Neural Information Processing Systems 30 (NIP 2017)