Modeling Wine Quality from Physicochemical Properties

Dale Angus
dangus@stanford.edu

Introduction

My experience of buying wine based on high scores is pretty good. One limitation that comes with my approach is that this greatly limits my choice of wines to the scored ones. I wish to be able to avail of good wines from vineyards that have no access to these expensive wine critics. So I am curious on how to score wines. An automated way, could there be?

I propose the development of **YAWDA**, the AI-powered wine-scoring system. The goal of YAWDA is to predict the wine rating that closely matches the one given by the humans.

YAWDA – Yet Another Wine Data Analysis

Data

The data I used is from the UCI repository. There are **two** datasets related to the **red** and **white** variants of the Portuguese "Vinho Verde" wine. The data only includes physicochemical (inputs) and sensory (output) variables. Data on grape types, wine brand and wine selling price are not included.

All input data were normalized using a Standard Scaler.

https://archive.ics.uci.edu/ml/datasets/Wine+Quality

Features

Input variables (based on physicochemical tests):

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
11 – alcohol

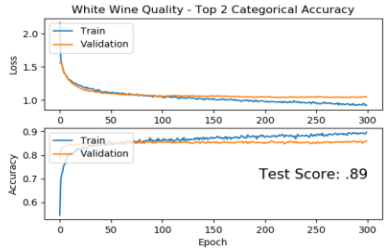
Output variable (based on sensory data):
12 - quality (score between 0 and 10)

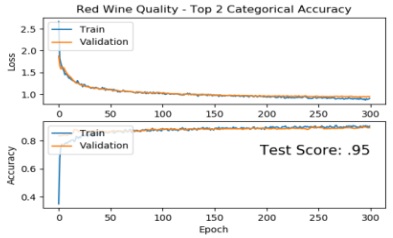
In the binary classifier model that uses the **combined** dataset, a label was added to indicate red (1) or white (0) variants.

Samples	Train	Validation	Test	Total
White	2742	1176	980	4898
Red	895	384	320	1599

Train-Validation-Test Split: 56-24-20

Predicting Wine Quality





All 11 features were needed to get the highest top-2 categorical accuracy.
Output: **probabilities for each score, top 2 checked**
Accuracy: **Red 95%**
Accuracy: **White 89%**

Comparison with Scikit-Learn Classifiers

Classifier / Top 2 Accuracy	White	Red
YAWDA (neural network)	0.89	0.95
ExtraTreesClassifier	0.92	0.94
VotingClassifier	0.91	0.93
GradientBoosting	0.89	0.89
XGBClassifier	0.87	0.93
MLPClassifier	0.88	0.90
GaussianProcessClassifier	0.87	0.94

Discussion

The goal of this project is to predict the quality rating that closely matches the actual rating. The metric **Top-k Categorical Accuracy** is the perfect metric to measure this goal. The model did a good job in achieving decent accuracy scores and is as good as the best classifiers in Scikit-Learn (see table). Two key steps were noted, 1) normalizing the data turned out to be very important and 2) regularization using Dropout helped in preventing overfitting during training. Feature selection was done but the combination of all 11 features produced the best result.

Models

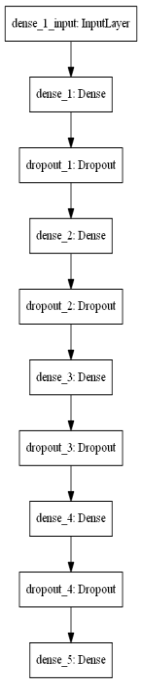
The models developed are **Deep Learning**-based model using a **6-layer neural network**. The models were developed using **Keras** with **Tensorflow**.

Multi-Class Classifier – Allowed two guesses, 1st and 2nd top predictions, can YAWDA correctly predict the quality?
Hidden layer activation: ReLU
Output layer activation: **Softmax**
Regularization: Dropout
Optimizer: Adam
Initializer: He Normal
Loss function: categorical cross-entropy

$$-\sum_{c=1}^{M} y_{o,c} \log(p_{o,c})$$

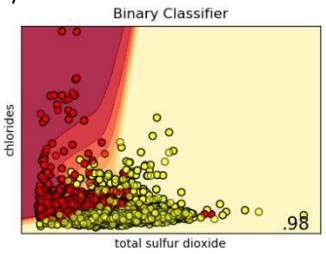
Binary Classifier – Can YAWDA predict the variant of wine, red or white?
Hidden layer activation: ReLU
Output layer activation: **Sigmoid**
Regularization: Dropout
Optimizer: Adam
Initializer: He Normal
Loss function: binary cross-entropy

$$-(y \log(p) + (1-y) \log(1-p))$$



Predicting Wine Variant

Model only needed two features to predict the variant of wine.
Input variables: **chlorides, total sulfur dioxide**
Output: **red or white**
Accuracy: **98%**



Model can achieve 99% accuracy when combined with other features, e.g., fixed acidity and pH.

Future

I would like to test if these models can be applied to other classification datasets and I would like to study regression analysis with other wine datasets. ☺