Proposal

September 30, 2019

1 Project description

As my final project I will atempt to create a model that predicts white wine quality based on its chemical properties. Hence, **the target variable is wine quality**.

The problem can be viewed as both regression or classification depending whether we view the wine quality as a continuous or discrete feature. We will approach the problem as a **classification** task because usually people use a discrete scale to describe how much they like something.

The problem is interesting because it attempts to investigate the **objective characteristics** of wine that people enjoy. If it turns out that perceived wine quality can be predicted from its chemical properties, the results might be impactful for wine producers (who will gain some insight how to produce better wines) and also for consumers (wishing to maximize their wine satisfaction). On the other hand, if it is the case that people's wine preference are independent from it's chemical properties then it would be an interesting result from a psychological perspective.

2 Data set

The dataset used in this proeject comes from the UCL Machine Learning Repository. It contains **4898 data points** and **12 features** and it is rather well documented, please see the winequality.names file for the original data set descrption. All features are numerical but they lack units of measurment. From the ranges of these feature we can infer the following units:

Feature	Units
fixed acidity	standard ph unit
volatile acidity	g/L
citric acid	g/L
residual sugar	g/L
chlorides	g/L
free sulfur dioxide	ppm
total sulfur dioxide	ppm
density	g/L
sulphates	g/L
pH	standard ph unit
alcohol	% by volume
quality	1-12 scale, 12 is the best

This dataset has been used in multiple public studies including the following three:

- *Modeling wine preferences by data mining from physicochemical properties* - Academic study that used regression and support vector machine algorithms to predict the wine quality from the given features. The authors reported that the use of SVM outperformed regression and neural networks.
- *Project Report:-Red Wine Quality Analysis Final* - This study uses analogical data set for red wine with the same features and comparable sample size. The author's explonatory data analysis employs regression on some engineered features.
- *Exploratory Data Analysis on Red Wine Quality* - This study is also uses the red wine data set from the example above. The author's data analysis focuses on use of stastistical techniques, and it does not use regression or SVMs in order to predict the wine quality.

3 Preprocessing

Since data is already in tidy format, it has only numerical features and no missing values we only need to apply a suitable scalling to each feature. Since all features have rather small and predictable range, we can apply min-max scaler to all of them. So in the preprocessed we have 12 features, including the target feature.

4 Github

Please find all the relevant files in this repository.