Introduction to Applied Machine Learning in R - SMU

Modeling wine preferences by data mining and physicochemical properties

Winter 2021

Final Project

In this project, we will use two data sets related to wine quality and their classification as red and white variants of the Portuguese "Vinho Verde" wine. Vinho verde is a unique product from the Minho (northwest) region of Portugal. Medium in alcohol, is it particularly appreciated due to its freshness (specially in the summer). More details can be found at: http://www.vinhoverde.pt/en/.

The data is available at UCI. These datasets are made publicly available for research purposes by 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. The data can be found at this link as a zip file (winequality.zip).

This project has the following objectives:

- In the first part we will try to predict the wine quality by using physicochemical properties.
- In the second part, we will try to build a classification model for the same purpose.
- In the third section you will predict if the wine is red or white.

Please submit both **rmd and html** files by **April 21, 5pm** using Dropbox.

Here is the grading scheme:

- 10% for the format in Rmarkdown: you should have no spelling errors; each section should be well titled, the messages and warnings in code chucks are controlled, large prints-outs of scripts are avoided, and each action/step is justified verbally when needed.
- 10% Preprocessing: tabulating the data proper way, checking NA's, group balances, etc.
- 80% Codes and their presentations.

You can seek help from others or check the Internet for assistance. However, please note that, the work that you submit should be only your own work. Any violation of trust on this principle will be considered as cheating.

Part 1: Downloading data and preprocessing (10 points)

1. Use `read.csv` and load both datasets (white and red) from the website and check their `str`.
2. Apply all necessary preprocesing steps.

Part 2: Predicting the wine quality (50 points)

We will use the dataset with only white wines. In this part we will build a predictive model by trying multiple alternatives to see if we can predict the wine quality through physicochemical properties.

A. Regression problem (20 points)

1. Consider the outcome variable `quality` as a continuous numerical variable. Use parametric and non-parametric methods and select the one that has the best predictive power. In using parametric models, you can use non-trainable ones, where you can also introduce polynomials and interactions, if you prefer.
2. Justify your pick by RMSPE and its uncertainty.
3. Identify the most important predictors.
4. Report the final verdict by using the test set.

B. Classification problem (30 points)

Unlike the regression problem above, now consider yourself as a paid data analyst working at one of the big wine companies in Porto, Portugal. Hence, your boss needs to know, at the end, how your model can be applied to wines as a classifier. This process requires several steps. In the first step, as before, we identify a winning classifier among many alternative classifiers. Second, we need to use the tuned classifier (the winner) to report its performance metrics (AUC etc.) and the confusion table calculated by the optimal threshold. Therefore, we still need to calculate the the optimal threshold by the J-index.

1. Now you will regroup the outcome variable, `quality`, into "less-than-average" and "better-than-average" by creating a binary variable.
2. Justify your pick by AUC and its uncertainty in both cases.
3. Identify the most important predictors.
4. Use the winner model and the test set, report AUC, uncertainty. What's the optimal threshold? What's its confusion table?

Part 3: Predicting the wine type, red or white. (40)

This is a classification problem. Append both datasets for white and red wines and create one dataset (make sure that you have an identifier for white and red types). Please report the results for the following steps:

1. The outcome variable is the type (red or white) you created. Use parametric and non-parametric classifiers and select the one that has the best predictive power.
2. Justify your choice of model by AUC and its uncertainty.
3. Identify the most important predictors.
4. What's the final results using the test data?