2019

# ST 557 Final Project

## CLASSIFICATION, CLUSTERING, PCA, MODEL EVALUATION

HA, NGOC

## Background

In this project I worked with 2 datasets of red and white wine samples. The datasets contain 11 attributes (pH, density…) and wine expert ratings 1599 red and 4989 white wines. There are 2 goals: to distinguish white wine from red wine; and to better understand which of these variables is most important to wine quality. Unless otherwise noted, all analyses in this project were performed on standardized data.

## Goal 1

### (1a) Is there a difference in mean vectors between red and white wines for these 11 chemical attributes? Which attributes seem to differ most between red and white wine?

For the first question, I performed a Hotelling's $T^2$ test to assess the equality of mean vectors between red and white wines. P-value is close to 0 and the null hypothesis is rejected, which means there is a significant difference in mean vectors between red and white wine.

To quantify which attributes differ the most between red and white wine, I performed a Linear Discriminant Analysis. The result is as follows:

| | fixed.acidity | volatile.acidity | citric.acid | residual.sugar | chlorides | free.sulfur.dioxide | total.sulfur.dioxide | density | pH | sulphates | alcohol |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LD1 | 0.4191934 | -0.5038349 | 0.127147 | 1.670847 | -0.1782635 | -0.3410196 | 1.134088 | -2.730096 | 0.1778424 | -0.1287368 | -0.9842167 |

Density and Residual Sugar have highest dispersion between red and white wine, followed by Total Sulfur Dioxide and Alcohol.

### (1b) Classification

I decided to start my classification analysis with a k-Nearest Neighbor (kNN) approach. If the error rate turned out to be high, I would then try to use a more sophisticated classifier (e.g. logistic regression, random forests, etc.). The reason for this approach is Occam's Razor: always use the simplest model that can make sufficiently accurate predictions.

The dataset is split into 3 subsets - training, validation, and testing – with 60:20:20 ratio. The validation set is used for finding a good *k* value to use for my kNN classifier. By performed a grid search with k ranging from 1 to 20, I chose k = 3, which gave the smallest Apparent Error Rate (AER) on the validation set. AER on the testing set is with k = 3 is: 0.00539707 = 0.54%. Since AER on testing set is very small, a more sophisticated classifier is unnecessary.

### (1c) Clustering

I started my clustering analysis with a k-means (k=2) approach, using Euclidean distance. If the performance from this approach was not desired, I would turn to Mahalanobis distance to use the data more efficiently (again, per Occam's Razor).

After obtaining 2 clusters from k-means, I assessed how well these clusters reflected the red/white classification, I calculated Gini Impurity of each cluster using the true labels (red and white). The result is as follows:

- Gini Impurity for Cluster 1: 0.07934953
- Gini Impurity for Cluster 2: 0.009839858

The clusters are very pure (small Gini Impurities) so Euclidean distance is good enough.

## Goal 2

### (2a) Is there a difference in mean vectors between wines with different quality scores? (Low, Medium, High)

I performed a MANOVA to assess the equality of mean vectors between wines of low, medium and high quality. P-value is close to 0 and the null hypothesis is rejected, which means there is a significant difference in mean vectors between wines of different quality.

### (2b) Regression

My initial classifier is a Multiple Linear Regression model. Its Mean Squared Error (MSE) on the test set is 0.411127805962143. Next, I used a Random Forests classifier and achieved a smaller MSE of 0.351242045 on the test set.

### (2c) Regression on PCA

The Random Forest classifier performed on the first 2 PCs gave an AER of 0.5097 on the test set. This means the first 2 PCs left out a lot of information of the original dataset, leading to poorer regression performance.

Appendix