ECE 532 Project Proposal: Desirable Characteristics of Wine
By: Danica Fliss

Project Summary

The objective of this project is to determine if there are certain characteristics of wine, or combinations of characteristics, that determine whether a wine is "quality" wine. To do this, we will utilize linear regression, principal component analysis (PCA), and a k-means algorithm. By the end of this project, we should be able to determine if particular characteristics are more important in wine quality, using linear regression and PCA, or if there is a magic combination of characteristics, using k-means algorithm and PCA.

Dataset

This project has a lot of applications outside of wine quality control, but for this project we will be using the Wine Quality Dataset from the University of California – Irvine's Center for Machine Learning and Intelligent Systems, which was originally used in research done by Cortez et al. [1], as an example. The Wine Quality Dataset consists of two sub-categories, red wines and white wines, of the vinho verde wine from northern Portugal. The red wine dataset has 1599 entries; the white wine dataset has 4898 entries. Each entry consists of 11 characteristics or "features" based on "physicochemical" attributes and one quality rating or "label" on the scale from 1 to 10 [1].

Outline

The first step is to analyze the red and white wines separately. To do this, we need to split each dataset into training data and testing data. The training data will be used to build the models, and the testing data will be used to determine how well the model generalizes to the problem. We will use a simple linear regression model to see how well the features define wine quality, and the squared error will serve as a baseline indicator on model accuracy.

Next, we will use the singular value decomposition (SVD) on the entire dataset to find the principal components. From the first principal component, we can determine which features are most necessary when classifying wine quality. The condition number, the ratio between the first singular value and the rest will reveal the strength of the first principal component. If the strength is weak, then it may be that wine quality is not based on a linear combination of its characteristics.

After we find the condition number, we will want to see if particular combinations of features, not individual features, can account for quality. We will use a k-means algorithm to cluster wines into categories that share similar features. Using cross-validation, we will find an appropriate k. We will incrementally increase the value of k until the variance changes by less than some threshold. We will then add the final cluster number as a feature of the wines. Going back to the first step, we will use a linear regression model to predict wine quality. If the squared error decreases by a significant amount, then particular combinations of features may be more important than individual features. Finding the condition number from the singular values for the new feature matrix will also help determine the importance of the combinations.

Timeline

Deliver proposal	October 22, 2020
Parsing data into usable format	1 week
Create linear regression model and calculate squared error	1-3 days
Determining the most important features in the classification problem	2-4 days
Program the cross-validated k-means algorithm	1-2 weeks
Deliver first update	November 17, 2020
Add cluster numbers to the dataset	2-4 days
Re-evaluate the squared error and feature strength of the linear regression model with new data	1-3 day
Deliver second update	December 1, 2020
First draft of final report	1 week
Final edits on final report	1-3 days
Deliver Final Report	December 12, 2020

Link

https://github.com/flissd/ECE532_Final_Project

Sources

[1] 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.