Madeline Hayes
ECE532 Final Project Proposal
22 October 2020

Investigating Classification of Wine Quality and Color from Physiochemical Properties

Background: Advanced characterization of wine can be achieved by combining traditional sensory assessments with physiochemical tests. These data are useful for assuring quality and safety, targeting growing consumer markets, and pricing. Though assessment of wine by sensory classification is a time-honored tradition (and an important cultural marker in many countries), the connection between taste and chemical properties is still not well understood. Accumulating large datasets of both the quality assessment (human sensory judgement) and physiochemical properties (e.g. alcohol content, density, sugar content) provides an opportunity to construct predictive models that will allow producers to rapidly characterize their product and target distribution to appropriate markets.

Dataset of Choice: Dataset: Wine Quality Dataset (https://archive.ics.uci.edu/ml/datasets/wine+quality)

This dataset is composed of 6497 samples with 13 attributes. Each sample is a variant of the Protugese Vinho Verde, with 1599 reds and 4898 whites. These data have 12 chemical attributes (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol content) and a quality score (human taste rating from 1 to 10). This wide variety of data makes this dataset ideal for a variety of machine learning applications.

Proposed Research Questions: Can we predict the quality of a wine based on its chemical properties? Can we determine whether a wine is red or white from its chemical properties?

Approach #1: Regression

In this approach I will aim to predict a drinker's rating of a wine from its chemical composition. This approach utilizes a weighted sum of each feature (physiochemical property) to approximate the label (rating) for a sample. This is a supervised learning problem and can be optimized using cross validation. I hypothesize that the models for red and white wines will be different in their most significantly contributing features.

Approach #2: K-nearest neighbor (KNN)

In this approach I aim to predict the approximate rating of a wine from its chemical properties by choosing target ratings and classifying each wine based on its proximity to that rating. KNN is a supervised learning problem where the labels are provided. KNN can be optimized by k-folds cross validation. I hypothesize that this method may be more accurate for predicting labels than regression

Approach #3: K-means clustering

In this approach I am to classify a wine's color based on its chemical properties. This is an unsupervised problem where the algorithm will use the distances between points to determine cluster centers. In this problem I propose to combine the red and white datasets and compare the distributions of red and whites over a k=2 clustering problem. I hypothesize that excluding some features of the data will lead to better clustering, as the distribution of certain properties (e.g. alcohol content) is less likely to vary between reds and whites.

Proposed Timeline:

October 22 – Proposal submitted

November 17 – Update #1 – Progress will be the initial exploration of the dataset, examining distributions of each feature. Aim to have K-means clustering complete and KNN in progress.

December 2 – Update #2 – Progress will be completed KNN and regression modelling in progress. Final report write up in draft stage.

December 12 – Final Project Submitted

Github Repository: https://github.com/mmhayes/ECE532_Final