Madeline Hayes
ECE532 Final Project Update
17 November 2020

Investigating Classification of Wine Quality and Color from Physiochemical Properties

I. Initial Dataset Exploration

Each attribute of the samples was visually assessed by histogram and violin plot to visualize the shape and distribution of each attribute. Assessment of the data in this way can help determine if data values need to be scaled.

Code and Figures: See 'exploratory_stats.ipny' and 'exploratory_stats.pdf'

II. K-means Clustering

From Proposal: 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.

Update: K-means clustering with the full dataset and a trimmed dataset (where 2 features, quality rating and alcohol content have been removed) is complete. I determined which features to remove by examining the distributions of each feature; quality and alcohol content had the most similar spread and shape between reds and whites, and I predicted that these features would therefore be less predictive. Data preprocessing was minimal, data was scaled between 0 and 1. The original hypothesis was correct; the full dataset had an error rate of 47.2% and the trimmed dataset had an error rate of 1.97%. Additionally, the trimmed dataset converged faster (5 iterations vs. 8 iterations), and had a lower sum of squared distances (410.3 vs. 699.5).

Code: See 'kmeans.ipny'

Timeline Update: Project is proceeding according to schedule, no modifications.

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