Madeline Hayes
ECE532 Final Project Update
1 December 2020

Investigating Classification of Wine Quality and Color from Physiochemical Properties

I. Initial Dataset Exploration - Complete

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

II. K-means Clustering - Complete

Code: See 'kmeans.ipynb'

III. K-Nearest Neighbors – In progress

From Proposal: 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.

Update: At this stage I'm exploring different options for the KNN algorithm. Because the quality rating data is discreet (all ratings are whole numbers between 1 and 10) and I want this feature to be predicted, I have tried several different options. I'm still working out how to visualize whether the predicted cluster is accurate. I might use this method to alternatively predict wine color and see if any of the KNN algorithm options (other than brute force distances, which is essentially k-means clustering) is more accurate at k=2 clustering than k-means clustering.

IV. Linear Regression – In progress

From Proposal: 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.

Update: I have completed an initial lasso-regularized model for this data, and it appears to be extremely predictive. However, the regularization parameter optimization has not proceeded as expected; for these data, it seems that there is little to no noise and the distance between predicted and actual ratings rapidly converges to zero even with a lambda that is relatively high. I may modify this approach to also predict color.

Code: see 'linearregression.ipynb'

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

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