Abstract:

In this assignment, we implemented the brute force K Nearest Neighbors (KNN) model to classify the quality of white wine. The data set was acquired from the University of California, Irvine (UCI) wine quality data set. Various numbers of neighbors, distance functions, and weight functions were tested in the assignment. The results showed that the brute force KNN with 11 neighbors and the Manhattan distance function and the inverse distance weight has the best performance (0.868) by evaluating the F1 scores.

Data Summary

The wine quality data set was downloaded from the Center for Machine Learning and Intelligent Systems at the UCI. It was developed and calculated by Cortez and his fellows in 2009 (reference). Two subsets of red and white Portuguese “Vinho Verde” wine were included. In this study, we employed the data set of white wine and applied the brute force KNN to predict its quality. The quality was evaluated by scores in the range between 0 and 10. To perform the binary classification, we converted the quantitative scores to binary data: good (1) and bad (0) with the threshold of 5. White wine with a score smaller than or equal to 5 was considered as bad quality, whereas white with a score larger than 5 was good quality. Except for the dependent variable, there were eleven response variables in the data sets, which are fixed acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. Table I shows the descriptive statistics of dependent and response variables.

The data set was pre-processed before inputting to the KNN model. Figure I is a scatter plot showing the relationship among all variables. It indicates the residual sugar and free sulfur dioxide have a small correlation to the quality and the citric acid has co-linearity with the pH and density. Therefore, we dropped residual sugar, free sulfur dioxide, and citric acid from the response variables, taking the rest eight as predictors. Additionally, because of the different units of every variable, we used feature scaling to reduce the skewness of variance in the predictors. The Gaussian standardization function is shown in equation 1