Statistical Learning and Analysis
Instructor: Prof. Varun Rai
Assignment 2
Due March 13 by 8:59am

Wine comes in various colors, tastes and quality. Although you won't get to taste any here, you can predict its quality using statistical learning techniques. In Questions 2 and 3 in this assignment, you will predict the quality of wine from data on its composition.

The dataset contains physicochemical variables sensory variables about Portuguese Vinho Verde wine. It is on Canvas and also available from the UCI machine learning repository, https://archive.ics.uci.edu/ml/datasets/wine+quality.

As with Assignment 1, you can use R or Python to respond, uploading either links or files to Canvas.

Question 1 (No dataset needed)
Suppose we collect data for a group of students in a statistics class, where X_1 = hours studied, X_2 = undergraduate GPA, and Y = whether students receive an A. We fit a logistic regression and produce estimated coefficients $\beta_0 = -6$, $\beta_1 = 0.05$, and $\beta_2 = 1$.

a) Estimate the probability that a student who studies for 40 hours and has an undergrad GPA of 3.5 gets an A in the class.
b) How many hours would the student in part (a) need to study in order to have a 50% chance of getting an A in the class?

Question 2 (Using the provided dataset)

(a) Produce some numerical and graphical summaries of the *Red Wine* data. What patterns do you see?
(b) Create a binary variable, final_quality, that contains a 1 if quality contains a value above its mean, and a 0 if quality contains a value below its mean. Use the full data set to perform a logistic regression with final_quality as the response and other variables as predictors (besides the original quality variable). Provide a summary of your obtained results. (use the summary function in R/statsmodel in python). Do any of the predictors appear to be statistically significant? If so, which ones?
(c) Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.
(d) Perform KNN on the full data set, with several values of K, in order to predict final_quality. What test errors do you obtain? Which value of K seems to perform the best on this data set?

Question 3 (Using the provided dataset)

a) Split the data into a training set (80%) and a test set (20%).
b) Perform LDA on the training data in order to predict final_quality using other variables as predictors. What is the test error of the model obtained?
c) Perform QDA on the training data in order to predict final_quality using other variables as predictors. What is the test error of the model obtained?

Code Book

1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol
12 – quality

Relevant Papers

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. Available at: [Web Link]