Statistical Learning and Analysis
Instructor: Prof. Varun Rai

Assignment 2

Wine comes in various color, test and quality. Although you won't be able to taste any here but can certainly try to predict its quality using statistical learning techniques. In this assignment, you will play with a wine dataset having its composition and predict its quality. You can either use R language or python (jupyter notebook) to do this assignment.

The provided dataset is related to variants of the Portuguese "Vinho Verde" wine. Some physicochemical (inputs) and sensory (the output) variables are about the wine is provided (e.g. there is no data about grape types, wine brand, wine selling price, etc.). This *dataset is also available from the UCI machine learning repository, https://archive.ics.uci.edu/ml/datasets/wine+quality.*

Question 1 (No Dataset)
Suppose we collect data for a group of students in a statistics class with variables X1 = hours studied, X2 = undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, $\beta 0$ (hat) = −6, $\beta 1$ (hat) = 0.05, $\beta 2$ (hat) = 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 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. Do there appear to be any patterns?
(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. (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

For more information, read [Cortez et al., 2009].
Input variables (based on physicochemical tests):

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

Output variable (based on sensory data):
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]