CS 484: Introduction to Machine Learning

Spring 2021 Assignment 5

The Center for Machine Learning and Intelligent Systems at the University of California, Irvine manages the Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php). We will use the WineQuality_Train.csv for training and the WineQuality_Test.csv for testing.

The categorical target variable is *quality_grp*. It has two categories, namely, 0 and 1. The input features are *alcohol*, *citric_acid*, *free_sulfur_dioxide*, *residual_sugar*, and *sulphates*. These five input features are considered interval variables.

You will use these two datasets for answering Questions 1 and 2.

Question 1 (50 Points)

We will apply the Adaptive Boosting technique for training a classification tree model. The model specifications are as follows.

- The Splitting Criterion is the Entropy
- The maximum tree depth is 5
- The initial random state value is 20210415 for classification tree and Boosting
- The maximum number of Boosting iterations is 50
- Stop the iteration if the classification accuracy on the Training data is greater than or equal to 0.9999999
- If the observed *quality_grp* is 1, then the absolute error is 1 – Prob(*quality_grp* = 1). Otherwise, the absolute error is Prob(*quality_grp* = 1).
- If an observation is correctly classified, then the weight is the absolute error. Otherwise, the weight is the absolute error plus 2.
- If Prob(*quality_grp* = 1) $\geq$ 0.2, then the predicted *quality_grp* is 1. Otherwise, the predicted *quality_grp* is 0.

a) (10 points) What is the Misclassification Rate of the classification tree on the Training data at Iteration 0 (i.e., when all the weights are one)?

   Ans: Misclassification Rate of the classification tree on the Training data at Iteration 0 is 0.16736309654717396.

b) (10 points) What is the Misclassification Rate of the classification tree on the Training data at Iteration 1?

   Ans: The Misclassification Rate of the classification tree on the Training data at Iteration 1 is 0.153702635410365.

c) (10 points) What is the Misclassification Rate of the classification tree on the Training data when the iteration converges, if any?

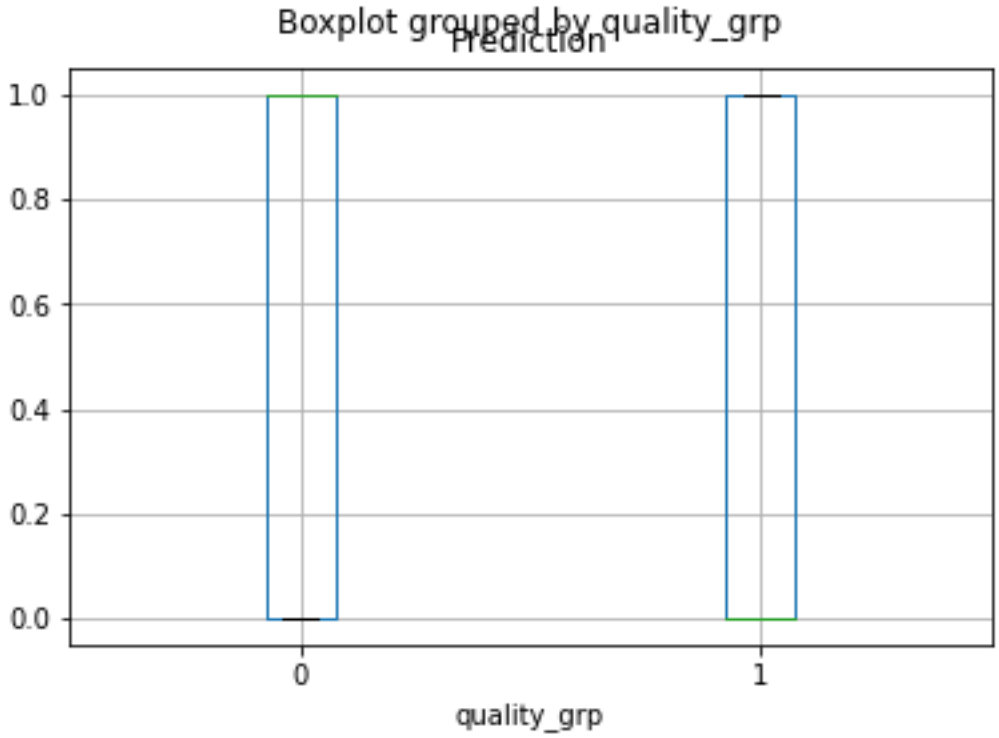
   Ans: The iteration is converged at Iteration 17 and the Misclassification Rate of the classification tree on the Training data is 0.0000001

d) (10 points) What is the Area Under Curve metric on the Testing data using the final converged classification tree?

   Ans: The Area Under Curve metric on the Testing data using the final converged classification tree is 0.3199273984283462

e) (10 points) Generate a grouped box-plot for the predicted probability for *quality_grp* = 1 on the Testing data. The groups are the observed *quality_grp* categories.

Ans:



Question 2 (50 points)

We often use the Area Under Curve metric to evaluate the goodness-of-fit of a binary classification model. Often, we need more than a point estimate to make our decisions. We want to train a logistic regression. We need your help to obtain the 95% confidence limits for the Area Under Curve metric on the Testing data.

a) (10 points) Use the Forward Selection method to select input features into the model. The final model must include the Intercept term. Use $\alpha = 0.05$. Which input features did you enter into the model?

Ans: The input features (selected using Forward Selection Method) shall be entered is

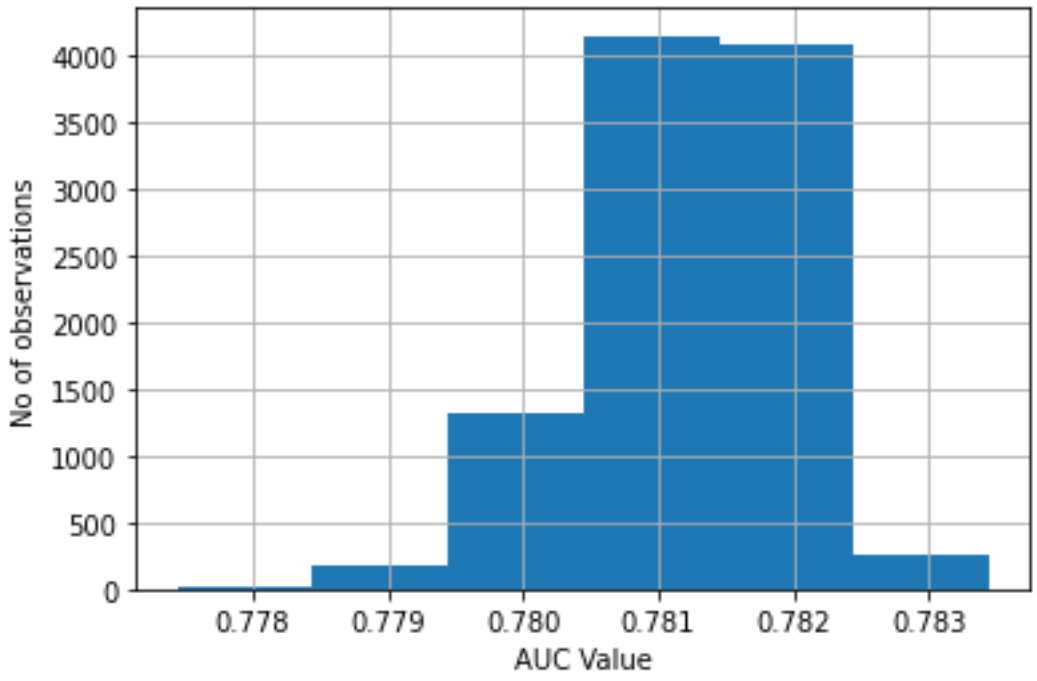
Model = Intercept + alcohol + citric_acid + free_sulfur_dioxide + residual_sugar + sulphates

b) (10 points) What is the Area Under Curve metric on the Testing data?

Ans: The Area Under Curve metric on the Testing data is 0.7817747664296713.

c) (10 points) Generate 10,000 Bootstrap samples from the Training data. Your random seed is 20210415. Then train a logistic regression model on each Bootstrap sample. The model will contain the input features that you have selected in (a). After each logistic regression model converges, calculate the predicted probabilities and the Area Under Curve metric on the Testing data. Generate a histogram of the 10,000 AUC metrics. The histogram width is 0.001.

Ans:



d) (10 points) Using the numpy.percentile function, calculate the 2.5th percentile and the 97.5th percentile of the 10,000 AUC metrics. What are the two percentile values?

Ans: 2.5th percentile is 0.7795629643034604 and 97.5th percentile is 0.7824446746578493 and 95% Confidence Interval is 0.7795630, 0.7824447.

e) (10 points) The two percentiles in d) will be the lower and the upper limits of the 95% confidence limits for the AUC on the Testing data. If the value 0.5 falls within the confidence limits, then statisticians will conclude that the AUC on the Testing data is not significantly different from 0.5. Based on your 95% confidence limits, what is your conclusion?

Ans:

The AUC on the testing data is significantly different from 0.5 and this does not fall in the intervals of confidence levels.