Tania Soutonglang

CS 484: Introduction to Machine Learning

Fall Semester 2022 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 20230101 for the 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) ≥ 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)?

Accuracy = 0.8326

Misclassification Rate = 0.1674

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

Accuracy = 0.8463

Misclassification Rate = 0.1537

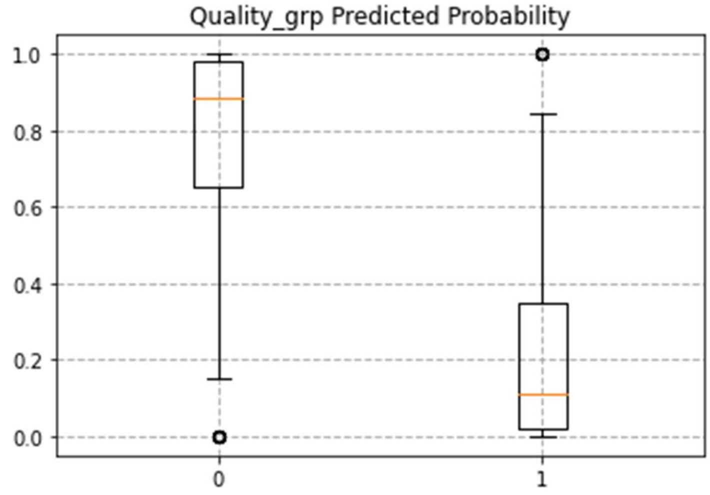
c) (10 points) What is the Misclassification Rate of the classification tree on the Training data when the iteration converges?

Accuracy = 1

Misclassification Rate = 2.4858e-08

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

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.



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 to go beyond 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.

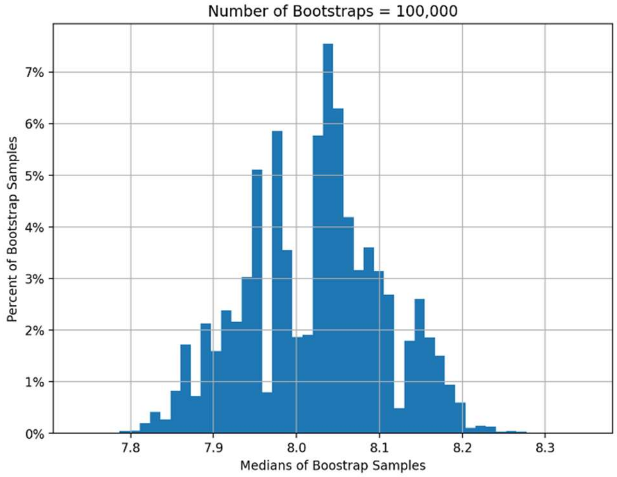
a) (10 points) Enter the five input features in this order: *alcohol*, *free_sulfur_dioxide*, *sulphates*, *citric_acid*, *residual_sugar*. The model must include the Intercept term. What is the Area Under Curve metric on the Training data?

0.7819

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

0.7724

c) (10 points) Generate 100,000 bootstrap samples from the Training data. Your random seed is 20221225. Then train the same logistic regression model in (a) on each Bootstrap sample. After each logistic regression model has converged, calculate the predicted probabilities and the Area Under Curve metric on the Testing data. Generate a histogram of the Area Under Curve metrics from the bootstrap samples. For your convenience, please use a histogram width of 0.001.



d) (10 points) Using the numpy.percentile function, calculate the 2.5th percentile, the 50th percentile, and the 97.5th percentile of the Area Under Curve metrics from the bootstrap samples. What are the percentile values?

7.8616, 8.0337, 8.1817

e) (10 points) The 2.5th and the 97.5th percentiles in (d) will be the lower and the upper limits of the 95% confidence limits for the Area Under Curve metrics on the Testing data. If the value 0.5 falls within the confidence limits, then statisticians will conclude that the Area Under Curve metric on the Testing data is not significantly different from 0.5. Based on your 95% confidence limits, what is your conclusion?

The AUC of the testing data is significantly different from 0.5.