CS 484: Introduction to Machine Learning

Spring 2021 Assignment 5 Answer Key

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) 0.2, then the predicted *quality\_grp* is 1. Otherwise, the predicted *quality\_grp* is 0.

1. (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)?

The accuracy is 0.7528040466241478. Thus, the misclassification rate is 0.2472.

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

At Iteration 1, the accuracy is 0.3334066417418. Thus, the misclassification rate is 0.6666.

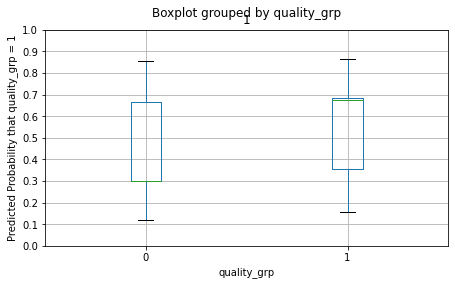
1. (10 points) What is the Misclassification Rate of the classification tree on the Training data when the iteration converges, if any?

The iteration did not converge because the accuracy was never higher than 0.9999999. In fact, the accuracy values bounce between 0.28832197053 and 0.7127776555971. At Iteration 49, i.e., the last iteration, the accuracy is 0.28832197053. Thus, the misclassification rate is 0.7117.

1. (10 points) What is the Area Under Curve metric on the Testing data using the final converged classification tree?

The Testing data have 1950 observations. The Area Under Curve on the Testing data is 0.8050.

1. (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.



The median for quality\_grp = 0 is 0.3027 and the first quartile is 0.2999. Therefore, the median line is practically indistinguishable form the lower boundary of the box. For your information, the summary statistics for the predicted probability for *quality\_grp* = 1 on the Testing data is:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| quality\_grp | Count | Mean | Std.Dev | Min. | Q1 | Median | Q3 | Max. |
| 0 | 1565 | 0.4011 | 0.1741 | 0.1195 | 0.2999 | 0.3027 | 0.6677 | 0.8583 |
| 1 | 385 | 0.5784 | 0.1667 | 0.1571 | 0.3566 | 0.6762 | 0.6833 | 0.8673 |

# 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.

1. (10 points) Use the Forward Selection method to select input features into the model. The final model must include the Intercept term. Use . Which input features did you enter into the model?

Below is the Step Summary table.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Step | Entered Predictor | Model | | Deviance | | |
| D.F. | Log-Likelihood | Chi-Squares | D.F. | Significance |
| 0 | const | 1 | -2250.9992 |  |  |  |
| 1 | alcohol | 2 | -1901.6280 | 698.7425 | 1 | 5.61E-154 |
| 2 | free\_sulfur\_dioxide | 3 | -1887.5393 | 28.1773 | 1 | 1.11E-07 |
| 3 | sulphates | 4 | -1878.8495 | 17.3797 | 1 | 3.06E-05 |
| 4 | citric\_acid | 5 | -1873.6313 | 10.4362 | 1 | 0.001235675 |
| 5 | residual\_sugar | 6 | -1869.8237 | 7.6153 | 1 | 0.005787443 |

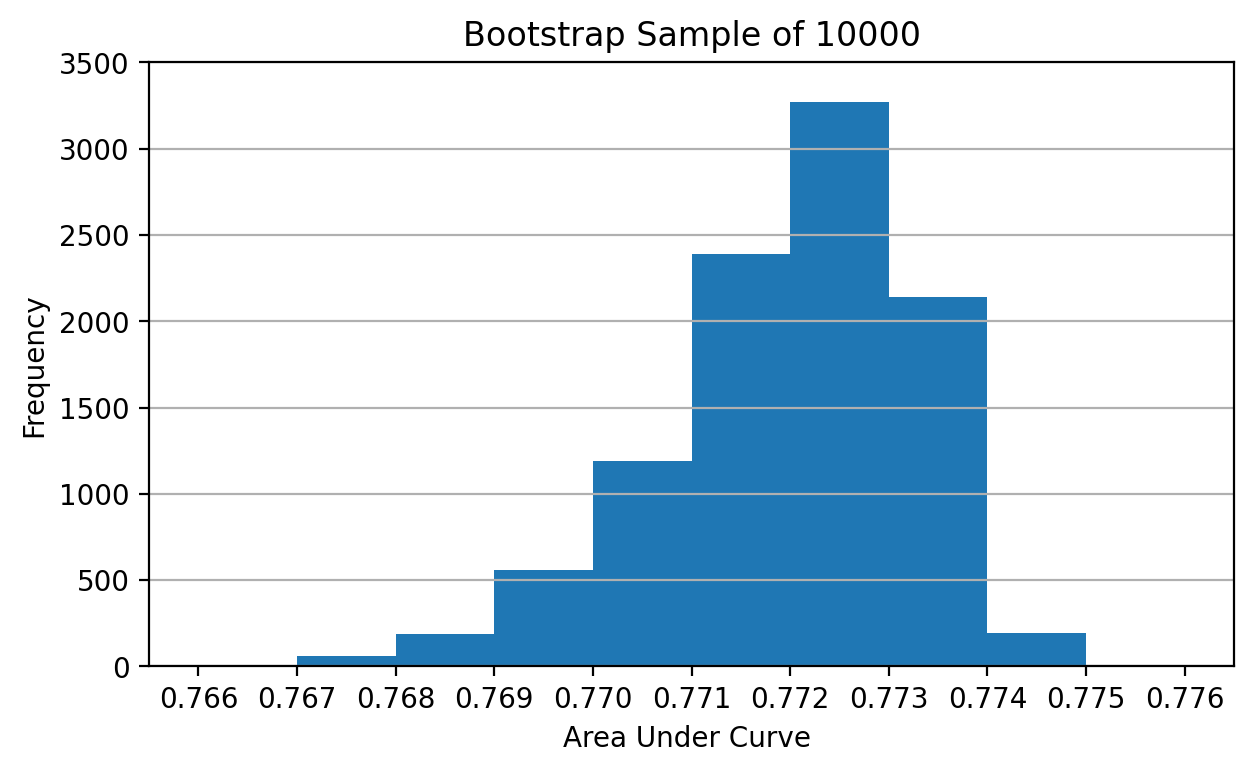
The input features enter into the model in this order: *alcohol*, *free\_sulfur\_dioxide*, *sulphates*, *citric\_acid*, and *residual\_sugar*.

1. (10 points) What is the Area Under Curve metric on the Testing data?

The Area Under Curve metric on the Testing data is 0.7724442969171403.

1. (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.

We trained one logistic regression model on each individual bootstrap sample. Then, we calculate the Area Under Curve metric on the Testing data. Eventually, we have 10,000 values of the metric. Below is the histogram.



1. (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?

The 2.5th percentile is 0.7689440 and the 97.5th percentile is 0.7739513.

1. (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?

The 95% confidence limits are (0.7689440, 0.7739513). Since the value 0.5 does not fall within the confidence limits, we will conclude that the AUC on the Testing data, which is 0.7724443, is significantly different from 0.5.