CS 484: Introduction to Machine Learning

Spring 2021 Assignment 3 Answer Key

# Question 1 (20 points)

An observation is misclassified if the predicted target category is different from the observed target category. The misclassification rate is the proportion of observations that are misclassified. The following diagram shows the classification tree for a binary target variable. The target categories are 0 and 1. Based on the diagram, please calculate the misclassification rate.

Diagram

Description automatically generated

The classification tree has five leaf (or terminal) nodes. Since the classification tree will classify all the observations in a leaf node to the majority target category in the leaf node, the observations that belong to other target categories will be misclassified. Therefore, we can calculate the misclassification rate by adding up the numbers of misclassified observations in the leaf nodes. The following table summarizes the calculation.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Node | Predicted Class | Number of Observations | | | |
| Subtotal | Class 0 | Class 1 | Misclassified |
| 0 | 0 | 119 | 119 | 0 | 0 |
| 1 | 1 | 1793 | 759 | 1034 | 759 |
| 2 | 0 | 1076 | 1076 | 0 | 0 |
| 3 | 0 | 3968 | 3117 | 851 | 851 |
| 4 | 0 | 1352 | 1351 | 0 | 0 |
| Total | | 8308 | 6422 | 1885 | 1610 |

The misclassification rate is 1610 / 8308 = 0.1938.

# Question 2 (40 points)

You will train a classification tree to predict the usage of a car. The data is the claim\_history.csv that contains 10,302 observations. The analysis specifications are:

**Target Field**

* **CAR\_USE**. The car’s usage. This field has two categories, namely, *Commercial* and *Private*.

**Nominal Feature**

* **CAR\_TYPE**. The car’s type. This feature has six categories, namely, *Minivan*, *Panel Truck*, *Pickup*, *SUV*, *Sports Car*, and *Van*.
* **OCCUPATION**. The occupation of the car owner. This feature has nine categories, namely, *Blue Collar*, *Clerical*, *Doctor*, *Home Maker*, *Lawyer*, *Manager*, *Professional*, *Student*, and *Unknown*.

**Ordinal Feature**

* **EDUCATION**. The education level of the car owner. This feature has five ordered categories which are *Below High School* < *High School* < *Bachelors* < *Masters* < *Doctors*.

**Decision Tree Specifications**

* Use only the complete records.
* The maximum number of branches is two.
* The maximum depth is two.
* The split criterion is the Entropy metric.

Since the sklearn tree module does not handle string features well, you should write custom Python codes to find the optimal split for a string feature. Also, do not encode the nominal features into dummy columns. It is because your classification tree is not deep enough to let all the dummy columns be used for splitting. Please answer the following questions.

1. (5 points). What is the entropy value of the root node?

The entropy value of the root node is 0.9489621.

1. (10 points). Please list the optimal split (i.e., feature name, values in the two branches, and the split entropy ) for all three features in the first layer.

The following table lists the optimal split for each predictor in the first layer. Please be reminded that the Split Entropy value will be the same even if we switch the Left and the Right Branches.

| **Predictor** | **Left Branch** | **Right Branch** | **Split Entropy** |
| --- | --- | --- | --- |
| CAR\_TYPE | ['Minivan', 'SUV', 'Sports Car'] | ['Panel Truck', 'Pickup', 'Van'] | 0.7684152 |
| OCCUPATION | ['Blue Collar', 'Student', 'Unknown'] | ['Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional'] | 0.7125833 |
| EDUCATION | ['Below High School'] | ['High School', 'Bachelors', 'Masters', 'Doctors'] | 0.9356143 |

1. (5 points). Which feature is selected for splitting in the first layer? What are the values in the branches of the first layer?

Among the three predictors, OCCUPATION has the lowest split entropy. Therefore, the split criterion of the first layer is OCCUPATION = ['Blue Collar', 'Student', 'Unknown'] and OCCUPATION = ['Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional'].

1. (10 points). Which features are selected for splitting in the second layer? What are the values in the branches of the second layer?

The following table lists the optimal split for each predictor in the second layer.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **First Layer** | **Predictor** | **Left Branch** | **Right Branch** | **Split Entropy** |
| OCCUPATION = ['Blue Collar', 'Student', 'Unknown'] | CAR\_TYPE | ['Minivan', 'SUV', 'Sports Car'] | ['Panel Truck', 'Pickup', 'Van'] | 0.7725783 |
| OCCUPATION | ['Student'] | ['Blue Collar', 'Unknown'] | 0.8042192 |
| EDUCATION | ['Below High School'] | ['High School', 'Bachelors', 'Masters', 'Doctors'] | 0.6670195 |
| OCCUPATION = ['Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional'] | CAR\_TYPE | ['Minivan', 'SUV', 'Sports Car'] | ['Panel Truck', 'Pickup', 'Van'] | 0.3274450 |
| OCCUPATION | ['Doctor', 'Home Maker', 'Lawyer'] | ['Clerical', 'Manager', 'Professional'] | 0.5664540 |
| EDUCATION | ['Below High School', 'High School', 'Bachelors'] | ['Masters', 'Doctors'] | 0.6175650 |

The selected predictors in the second layer are highlighted in light green.

1. (10 points). Describe the leaf (i.e., terminal) nodes in a table. Please include the decision rules, the counts of the target categories, and the predicted probabilities for CAR\_USE.

The following table describes the four leaf nodes of the decision tree.

| **Leaf** | **Predictor** | | | **CAR\_USE (Count / Probability)** | |
| --- | --- | --- | --- | --- | --- |
| **OCCUPATION** | **EDUCATION** | **CAR\_TYPE** | **Commercial** | **Private** |
| 0 | ['Blue Collar', 'Student', 'Unknown'] | ['Below High School'] |  | 216 (0.2625) | 607  (0.7375) |
| 1 | ['Blue Collar', 'Student', 'Unknown'] | ['High School', 'Bachelors', 'Masters', 'Doctors'] |  | 2559  (0.8448) | 470  (0.1552) |
| 2 | ['Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional'] |  | ['Minivan', 'SUV', 'Sports Car'] | 30  (0.0065) | 4564  (0.9935) |
| 3 | ['Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional'] |  | ['Panel Truck', 'Pickup', 'Van'] | 984  (0.5302) | 872  (0.4698) |

# Question 3 (40 points)

We provide you the sample\_v10.csv that contains 10,000 observations. This data contains a categorical variable **y** and ten continuous features are **x1**, **x2**, **x3**, **x4**, **x5**, **x6**, **x7**, **x8**, **x9**, and **x10**. You will then use this data to train a multinomial logistic regression model that always includes the Intercept term. To include only significant continuous features in the model, you will use the Backward Selection method to determine the list of significant continuous features. The threshold for test significance is 0.05.

1. (5 points). Show the frequency table of the categorical target field.

|  |  |  |  |
| --- | --- | --- | --- |
| **y** | 1 | 2 | 3 |
| Frequency | 2274 | 3532 | 4194 |

1. (5 points). What is the initial model in the Backward Selection method? Please also show the log-likelihood value and the number of free parameters.

The initial model in the Backward Selection model includes all the ten continuous features. The model form is logit(y) = Intercept + x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 + x10. The log-likelihood value is -1956.05 and the number of free parameters is 22.

1. (20 points). Please show the step summary of the Backward Selection method. The step summary should include the name of the removed feature, the log-likelihood value of the reduced model, the number of free parameters of the reduced model, the Deviance test statistic, the Deviance degree of freedom, and the Deviance significance value.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Step | Removed Predictor | Number of Free Parameters | Model Log-likelihood | Deviance Test | | |
| Chi-Square | DF | Significance |
| 0 |  | 22 | -1956.05 |  |  |  |
| 1 | x7 | 20 | -1956.07 | 0.0386 | 2 | 0.9809 |
| 2 | x3 | 18 | -1956.3 | 0.4558 | 2 | 0.7962 |
| 3 | x2 | 16 | -1956.58 | 0.5695 | 2 | 0.7522 |
| 4 | x5 | 14 | -1956.9 | 0.8246 | 2 | 0.6621 |
| 5 | x9 | 12 | -1957.9 | 1.8959 | 2 | 0.3875 |
| 6 | x6 | 10 | -1959.9 | 4.0197 | 2 | 0.1340 |
| 7 | x8 | 8 | -1961.9 | 3.8908 | 2 | 0.1429 |

1. (5 points). What is the final model suggested by the Backward Selection method?

The final model is logit(y) = Intercept + x1 + x4 + x10.

1. (5 points). Please calculate the Akaike Information Criterion and the Bayesian Information Criterion for all the models that you listed in (c). What model will each criterion suggest?

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Step | Removed Predictor | Number of Free Parameters | Model Log-likelihood | AIC | BIC |
| 0 |  | 22 | -1956.05 |  |  |
| 1 | x7 | 20 | -1956.07 | 3952.15 | 4096.36 |
| 2 | x3 | 18 | -1956.3 | 3948.6 | 4078.39 |
| 3 | x2 | 16 | -1956.58 | 3945.17 | 4060.54 |
| 4 | x5 | 14 | -1956.9 | 3941.99 | 4042.94 |
| 5 | x9 | 12 | -1957.9 | 3939.89 | 4026.42 |
| 6 | x6 | 10 | -1959.9 | 3939.91 | 4012.01 |
| 7 | x8 | 8 | -1961.9 | 3939.81 | 3997.49 |

The Akaike Information Criterion suggests this model Intercept + x1 + x4 + x6 + x8 + x10.

The Bayesian Information Criterion suggest this model Intercept + x1 + x4 + x10.