Tania Soutonglang

CS 484: Introduction to Machine Learning

Fall 2022 Assignment 3

Question 1 (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:

Label 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, you have to write your own Python codes to find the optimal split for a string feature. You must use values of a nominal string AS IS. Do not encode the nominal features into dummy columns. It is because your classification tree will not have the enough depth to allow all the dummy columns be used for splitting.

Please answer the following questions.

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

The entropy value of the root node is approximately 0.94896.

b) (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.

Feature Name	Left Branch	Right Branch	Split Entropy
CAR_USE	'Minivan', 'SUV', 'Sports Car'	'Panel Truck, 'Pickup, Van'	0.76842
OCCUPATION	'Blue Collar', 'Unknown', 'Student'	'Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional'	0.71258
EDUCATION	'High School'	'Below High School', 'Bachelors', 'Masters', 'Doctors'	0.93561

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

'OCCUPATION' has the lowest entropy of the three features so it will be used to split the first layer.

Branch 1 = ['Blue Collar', 'Unknown', 'Student']

Branch 2 = ['Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional']

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

'CAR_USE' and 'EDUCATION' will be selected to split the second layer.

Branch 1 = ['Minivan', 'SUV', 'Sports Car']

Branch 2 = ['Panel Truck, 'Pickup, 'Van']

Branch 3 = ['High School']

Branch 4 = ['Below High School', 'Bachelors', 'Masters', 'Doctors']

e) (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.

Leaf	OCCUPATION	EDUCATION	CAR_TYPE	Commercial	Private

0	['Blue Collar', 'Unknown', 'Student']	['Minivan', 'SUV', 'Sports Car']			
1	['Blue Collar', 'Unknown', 'Student']	['Panel Truck, 'Pickup, 'Van']			
2	['Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional']		['High School']		
3	['Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional']		['Below High School', 'Bachelors', 'Masters', 'Doctors']		

Question 2 (40 points)

We provide you the sample_v10.csv that contains 10,000 observations. This data contains a categorical label 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 Forward Selection method to determine the list of significant continuous features. The threshold for test significance is 0.05.

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

Target	0	1	2
Frequency	3529	2277	4194

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

The initial model was a multinominal logistic model.

log-likelihood value: -10689.33236

free parameters: 8

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

Step	Feature	Log-Likelihood Value	Free Parameters	Test Statistic	Degree of Freedom	Significance Value
0		-10,689.33236	2	NaN	NaN	NaN
1	x4	-8,235.400271	4	4,907.864172	2.0	0.0
2	x10	-2,250.764738	6	11,969.27106	2.0	0.0
3	x1	-1,985.566097	8	530.3972828	2.0	6.694119057e-116

d) (5 points). What is the final model suggested by the Forward Selection method?

An intercept only model

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

Step	Feature	Log-Likelihood Value	Free Parameters	Akaike Information Criterion	Bayesian Information Criterion
0		-10,689.33236	2	NaN	NaN
1	x4	-8,235.400271	4	-21,382.66471	-98,456.38935
2	x10	-2,250.764738	6	-16,478.80054	-75,858.83959
3	x1	-1,985.566097	8	-4,513.529477	-20,742.30934

Question 3 (20 points)

An observation is misclassified if the predicted target category is not the same as the observed target category. The misclassification rate is the proportion of observations that have been misclassified. The following diagram shows the classification tree for a binary target variable. The target categories are 0 and 1. Based on the diagram, what is the misclassification rate?