CS 484: Introduction to Machine Learning

Michal Malek

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 of the root node is .949

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.

The split entropy is 0.714 looking at student, blue collar and unknown from occupation.

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

In the first layer, we split for car type on the right side and education on the lefts side

Car type has values minivan, SUV and sports car. On the left side, education has a value for below high school.

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

Splitting at this layer, splits into our leaf nodes where we use private and commercial car use to split

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.

Left Left	Left Right	Right Left	Right Right
Split based on Below high school	Is split based on education and this one contains all except below high school	Split based on van, suv, sports car	Is split based on car type, everything but what's in right left

The decision rules come from the predicted probability from each observation. The predicted probabilities I got are 24.64%, 85.04%, 0.61% and 54.64%.

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.

index	y
1	2277
0	3529
2	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.

	Predictor	Type	ModelDF	ModelLLK
0	Intercept		2	-10689.332

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.

	Predictor	Type	ModelDF	ModelLLK	DevChiSq	DevDF	DevSig
0	Intercept		2	-10689.332			
1	x4	interval	4	-8235.400	4907.864	2.0	0.0
2	x10	interval	6	-2250.764	11969.271	2.0	0.0
3	x1	interval	8	-1985.566	530.397	2.0	6.694e-116

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

After 3 steps, I get the following model

'x8'	'interval'	10	-1983.396	4.340	2	0.114
'x6'	'interval'	10	-1983.565	4.001	2	0.135
'x9'	'interval'	10	-1984.704	1.722	2	0.422
'x5'	'interval'	10	-1985.100	0.931	2	0.627
'x2'	'interval'	10	-1985.307	0.517	2	0.772
'x3'	'interval'	10	-1985.399	0.333	2	0.846
'x7'	'interval'	10	-1985.530	0.072	2	0.964

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?

AIC = 21394.6

BIC = 21452.282

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?

Based on the diagram, when you take the total number of samples, which is the total number of outcomes and divide that by the total number of outcomes that are NOT equal to the predicted outcome, we get our answer for the misclassification rate. Looking at the root node, which contains all our samples and the number of outcomes that weren't like our predicted outcome, we get the following equation which is our misclassification rate.

$$\frac{1885 \quad = \textit{number of samples NOT equal to predicted outcome}}{8307 \quad = \textit{total numebr of samples}.}$$

$$= 0.227$$

= 22.7%