CS 484: Introduction to Machine Learning

Spring Semester 2023 Assignment 3

We provide you with the **claim\_history.xlsx** which contains 10,302 observations on various vehicles. You will use the observations in this Excel file to train models that predict the usage of a vehicle. Your models will use the following variables.

**Label Field**

* **CAR\_USE**. Vehicle Usage. It has two categories, namely, *Commercial* and *Private*.

**Nominal Predictor**

* **CAR\_TYPE**. Vehicle Type. It has six categories, namely, *Minivan*, *Panel Truck*, *Pickup*, *SUV*, *Sports Car*, and *Van*.
* **OCCUPATION**. Occupation of Vehicle Owner. It has nine categories, namely, *Blue Collar*, *Clerical*, *Doctor*, *Home Maker*, *Lawyer*, *Manager*, *Professional*, *Student*, and *Unknown*.

**Ordinal Predictor**

* **EDUCATION**. Highest Education Level of Vehicle Owner. It has five ordered categories which are *Below High School* < *High School* < *Bachelors* < *Masters* < *Doctors*.

Although a decision tree can accommodate missing values in the predictors, we will use only observations where there are no missing values in all the above four variables. After dropping the missing values, we will use all the 100% complete observations for training both models.

For each observation, you will calculate the predicted probabilities for CAR\_USE = *Commercial* and CAR\_USE = *Private*. You will classify the observation in the CAR\_USE category that has the highest predicted probability. In case of ties, choose the *Private* category.

# Question 1 (50 points)

You will train a classification tree model with the following specifications:

* The maximum depth is two.
* The split criterion is the Entropy metric.
* An observation in the parent node will be assigned to the left child node if the splitting criterion is evaluated to be True. Otherwise, it will be assigned to the right child node.

*Since the sklearn tree module does not handle string features, you have to write 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. To find all the possible splits of a nominal predictor, we suggest the itertools.combinations() function to you.*

1. (20 points) Please describe the leaf nodes of the classification tree. Your description should include these five pieces of information: (1) Splitting Criterion, (2) Number of Observations, (3) Predicted Probabilities of CAR\_USE, (4) Predicted CAR\_USE category, and (5) Split Entropy Value.
2. (10 points) Let us study a fictitious person. The person works in a *Professional* occupation, has an education level of *Doctors*, and owns a *Minivan*. What are the Car Usage probabilities?
3. (10 points) Let us study another fictitious person. The person is a *Student*, has a *Below High School* level of education, and owns a *Sports Car*. What are the Car Usage probabilities?
4. (5 points) Generate a histogram of the predicted probabilities of CAR\_USE = *Private*. The bin width is 0.05. The vertical axis is the proportion of observations.
5. (5 points) Finally, what is the misclassification rate of the Classification Tree model?

# Question 2 (50 points)

You will train a Naïve Bayes model. You will apply the Laplace/Lidstone value of 0.01 to the cell counts for the purpose of computing the row probabilities. However, you do not change the cell counts.

1. (10 points) What are the Class Probabilities?
2. (10 points) Cross-tabulate the label variable by each predictor and show the resulting table. The table must contain the frequency counts and the row probabilities in each label class.
3. (10 points) Let us study a fictitious person. The person works in a *Professional* occupation, has an education level of *Doctors*, and owns a *Minivan*. What are the Car Usage probabilities?
4. (10 points) Let us study another fictitious person. The person is a *Student*, has a *Below High School* level of education, and owns a *Sports Car*. What are the Car Usage probabilities?
5. (5 points) Generate a histogram of the predicted probabilities of CAR\_USE = *Private*. The bin width is 0.05. The vertical axis is the proportion of observations.
6. (5 points) Finally, what is the misclassification rate of the Naïve Bayes model?