CS 484: Introduction to Machine Learning

Autumn 2020 Assignment 3

Question 1 (10 points)

Prove that $E = -\sum_{j=1}^{K} p_j \log_2(p_j)$ attains its maximum value when $p_j = p_K = 1/K$.

Hint: (1) re-express $E = -\sum_{j=1}^{K-1} p_j \log_2(p_j) - p_K \log_2(p_K)$, (2) use this equality $\sum_{j=1}^{K} p_j = 1$ in calculating the partial derivatives $\partial E/\partial p_j$, $j = 1, \dots, (K-1)$, and (3) solve the equations $\partial E/\partial p_j = 0$, $j = 1, \dots, (K-1)$.

$$E = -\sum_{j=1}^{K} p_j \log_2(p_j) \rightarrow E = -\sum_{j=1}^{K-1} p_j \log_2(p_j) - p_K \log_2(p_K)$$

$$\partial E/\partial p_j = -\partial/\partial p_j \sum_{j=1}^{K-1} p_j \log_2(p_j) - p_K \log_2(p_K)$$

$$= -\sum_{j=1}^{K-1} \partial/\partial p_j \,(p_j \log_2(p_j) - p_K \log_2(p_K))$$

$$= -\sum_{j=1}^{K-1} \log_2(p_j) + p_j \frac{1}{p_j \ln(2)} = -\sum_{j=1}^{K-1} \log_2(p_j) + \frac{1}{\ln(2)}$$

$$= -\sum_{j=1}^{K-1} \log_2(p_j) - \sum_{j=1}^{K-1} \frac{1}{\ln(2)}$$

$$\partial E/\partial p_j = -\frac{K-1}{\ln\,(2)} - \sum_{j=1}^{K-1} \log_2(p_j)$$

$$= \cdots$$

For entropy to attain its maximum value, the distribution of the target values should be uniform. This means that they are completely impure and the probability of each are the same.

Thus, with $E = -\sum_{j=1}^{K} p_j \log_2(p_j)$ attains its maximum value when $p_j = 1/K$

Question 2 (10 points)

Suppose the predicted target value of a new observation is randomly assigned one of the target categories according to the categories' distribution. Argue analytically that the Gini Impurity is the probability of incorrect classification.

Hint: Probability of incorrect classification is the sum of these products of probabilities Prob(Do Not Classify to Category i given the observation is from the Category i) × Prob(an observation is drawn from the Category i).

In a pure node, target categories are the same so any observations cannot be misclassified, so probability of misclassification is 0. However, in a completely impure node, target categories are uniformly distributed, so any observation will have a 1/K probability of being correctly classified. Thus, the probability of a misclassification is 1 – 1/K.

With uniform distribution, $\boldsymbol{p_{ij} = \frac{1}{k}}$ **then** $\boldsymbol{Gini\ impurity = 1 - \sum_{j=1}^{k} p_{ij}^2 = 1 - \sum_{j=1}^{k} \frac{1}{k^2} = 1 - \frac{1}{k}}$

We see that these values are the same, thus, Gini impurity is the probability of incorrect classification.

Question 3 (10 points)

Argue analytically that a completely impure node yields the highest Gini Impurity.

Gini impurity is the probability of misclassification. In a completely impure node, target categories are uniformly distributed, so any randomly chosen observation will have a 1/K probability of being correctly classified. Thus, the probability of a misclassification is 1 – 1/K. You can see that when node is impure, the more target categories, the higher the probability of misclassification. So, uniform distribution, or a completely impure node, will result in the highest probability of misclassification and thus, highest Gini impurity. In conclusion, a completely impure node yields the highest Gini Impurity.

Question 4, 5, and 6

You will train a decision tree model 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 usage of car. This field has two categories, namely, *Commercial* and *Private*. The *Commercial* category is the Event value.

Nominal Feature

- **CAR_TYPE**. The type of car. This feature has six categories, namely, *Minivan*, *Panel Truck*, *Pickup*, *SUV*, *Sports Car*, and *Van*.
- **OCCUPATION**. The occupation of 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 car owner. This feature has five ordered categories which are *Below High School* < *High School* < *Bachelors* < *Masters* < *Doctors*.

Analysis Specifications

- **Partition**. Specify the target field as the stratum variable. Use stratified simple random sampling to assign 70% of the observations to the Training partition, and the remaining 30% of the observations into the Test partition. The random state is 60616.
- **Decision Tree**. The maximum number of branches is two. The maximum depth is two. The split criterion is the Entropy metric.

Question 4 (10 points)

Please answer the following questions about your Data Partition step. You may call the `train_test_split()` function in the `sklearn.model_selection` module in your code.

a) (5 points). Please provide the frequency table (i.e., counts and proportions) of the target field in the Training partition?

Commercial: 2652, 0.367

Private: 4559, 0.632

b) (5 points). What is the probability that an observation will be assigned to the Test partition given that CAR_USE is *Private*?

Probability: 0.2999

Question 5 (30 points)

Please provide information about your decision tree that is trained on the Training partition. You will need to write your own Python program to find the answers.

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

0.9489

b) (5 points). What is the split criterion (i.e., feature name and values in the two branches) of the first layer?

With entropy of 0.719, Occupation is the split criterion of the first layer with the values in the branches being {Clerical, Doctor, Home Maker, Lawyer, Manager, Professional} and {Blue Collar, Student, Unknown}

c) (5 points). What is the entropy of the split of the first layer?

The entropy of the split of the first layer is 0.719. After the split, the left branch results in a node(Education) with entropy of 0.679 and right branch results in node(Car Type) with

entropy of 0.335. The education node has branches of (Below High School) and (High School, Bachelors, Master, Doctors). The car type node has branches of (Minivan, SUV, Sports Car) and (Panel Truck, Pickup, Van)

d) (5 points). Describe all your leaves (i.e., terminal nodes) in a table. Please include the decision rules and the counts of the target values.

Leave 1

('Blue Collar', 'Student', 'Unknown') -> (Below High School)

Total Count: 578

Commercial: 155

Private: 423

Leave 2

('Blue Collar', 'Student', 'Unknown') -> (High School, Bachelors, Masters, Doctors)

Total Count: 2108

Commercial: 1770

Private: 338

Leave 3

('Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional') -> ('Minivan', 'SUV', 'Sports Car')

Total Count: 3225

Commercial: 27

Private: 3198

Leave 4

('Clerical', 'Doctor', 'Home Maker', 'Lawyer', 'Manager', 'Professional') -> ('Panel Truck', 'Pickup', 'Van')

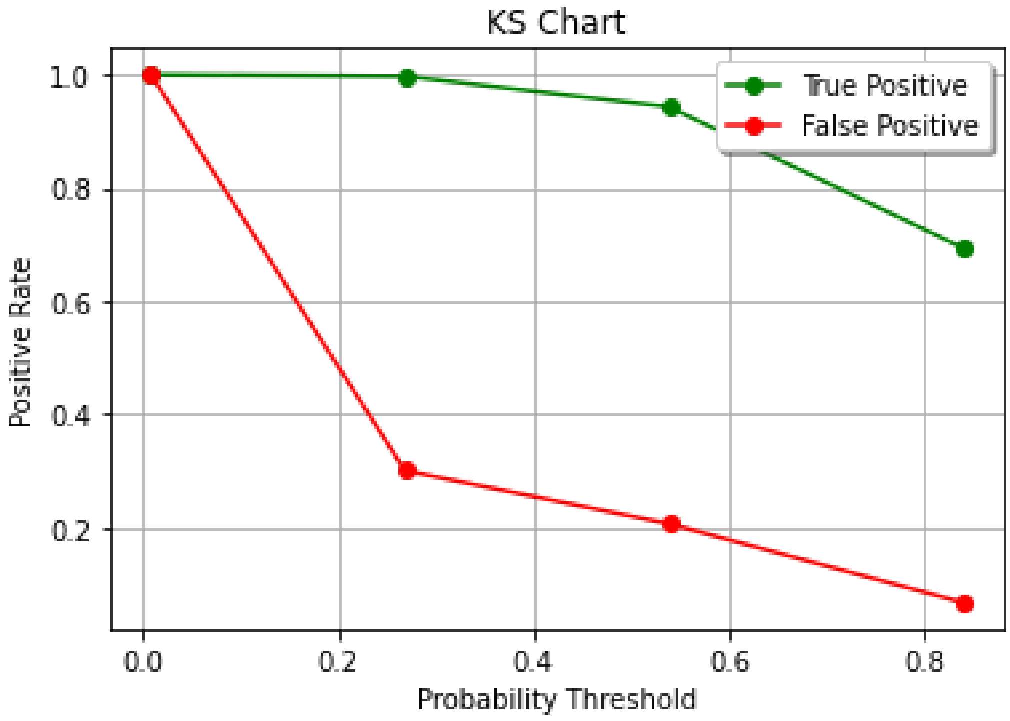
Total Count: 1300

Commercial: 700

Private: 600

e) (5 points). What is the Kolmogorov-Smirnov statistic?

0.737



f) (5 points). What is your suggested event probability cutoff value?

0.538

Question 6 (30 points)

Please apply your decision tree to the Test partition and then provide the following information. You will choose whether to call sklearn functions or write your own Python program to find the answers.

a) (5 points). Based on your suggested Kolmogorov-Smirnov event probability cutoff value as the threshold, what is the Misclassification Rate in the Test partition?

0.155

b) (5 points). What is the Root Average Squared Error in the Test partition?

0.312

c) (5 points). What is the Area Under Curve in the Test partition?

0.927

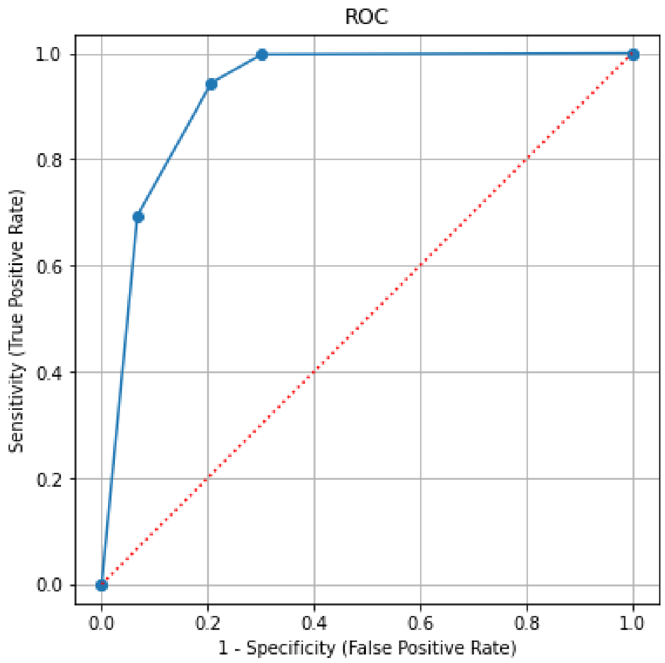
d) (5 points). What is the Gini Coefficient in the Test partition?

0.854

e) (5 points). What is the Goodman-Kruskal Gamma statistic in the Test partition?

0.937

f) (5 points). Generate the Receiver Operating Characteristic curve for the Test partition. The axes must be properly labeled. Also, include the diagonal reference line.


ROC
Sensitivity (True Positive Rate)
1 - Specificity (False Positive Rate)