CS 584 Machine Learning

Autumn 2019 Final Examination

# Instruction

1. Unless otherwise, please round your answer to the fourth decimal place.

# Question 1 (5 points)

The CSV file FinalQ1.csv contains 100 values. When we construct a histogram on these 100 values with a bin-width of 2, what is the empirical density value for x = 100?

# Question 2 (5 points)

Suppose the itemset {X, Y, Z} has 100% Support, then what is the Lift value of the rule {Y} 🡺 {X, Z}?

Multiple Choice:

1. Less than 1
2. Equal to 1
3. Greater than 1
4. Cannot be determined

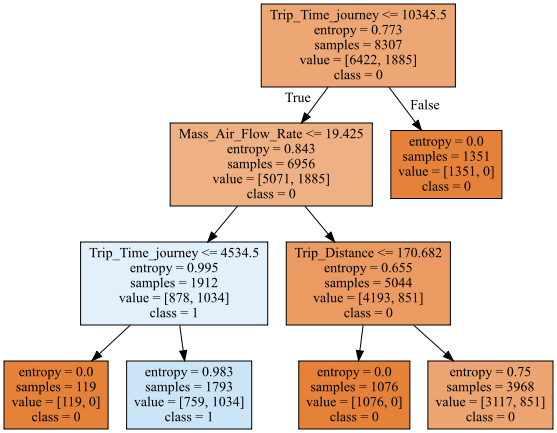
# Question 3 (5 points)

I calculated the Elbow values and the Silhouette values for my 1-cluster to 10-cluster solutions. Based on these values, what do you suggest for the number of clusters? An integer answer is expected.

|  |  |  |
| --- | --- | --- |
| Number of Clusters | Elbow Value | Silhouette |
| 1 | 579857.9543 | N/A |
| 2 | 532455.2722 | 0.5391 |
| 3 | 493218.0813 | 0.5300 |
| 4 | 433215.8150 | 0.5479 |
| 5 | 430290.4574 | 0.5411 |
| 6 | 412804.9312 | 0.5140 |
| 7 | 409729.7423 | 0.5172 |
| 8 | 404285.7518 | 0.5081 |
| 9 | 378087.1355 | 0.5056 |
| 10 | 369686.6227 | 0.4984 |

# Question 4 (5 points)

You are given the following classification tree diagram. The target variable has two categories 0 and 1.



We are going to ignore the predicted class suggested in the tree diagram. Instead, our classification rule says that if Prob (class = 1) ≥ 0.2, then the predicted Class is 1. Otherwise, the predicted class is 0. What is the misclassification rate?

# Question 5 (5 points)

The CSV file FinalQ5.csv contains 500 observations. It has three columns.

1. TransportMode: mode of transportation, a categorical predictor that has four levels: *Bike*, *Drive*, *Public*, and *Walk*
2. CommuteMile: number of miles for commuting, an interval predictor
3. Late4Work: late for work indicator, a binary target that has two levels: *No* and *Yes*

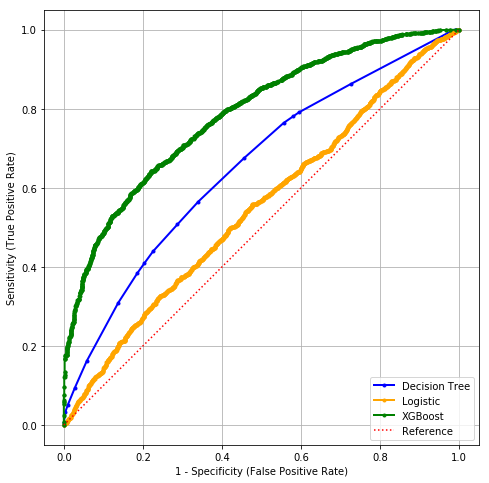
You will build a main effect logistic regression model that includes the Intercept term. Based on the model results, what phenomenon can you conclude the model is having?

Multiple Choice:

1. No Special Phenomenon
2. Complete Separation
3. Quasi-Complete Separation in One Combination of TransportMode and CommuteMile
4. Quasi-Complete Separation in Two Combinations of TransportMode and CommuteMile
5. Quasi-Complete Separation in Three Combinations of TransportMode and CommuteMile

# Question 6 (5 points)

Three models, namely, the logistic, the decision tree, and the XGBoost, are built on 5,155 observations. The Receiver Operating Characteristics curves are shown below.



If I need at least 4,124 True Positive observations, then how much more False Positive observations do I have to tolerate if I choose the Decision Tree model over the XGBoost?

Multiple Choice:

1. Approximately 3,100
2. Approximately 2,100
3. Approximately 1,000
4. Approximately 1,800
5. Approximately 770

# Question 7 (5 points)

You will calculate the Eta-squared statistic to measure the association between the interval target MPG\_Highway and the categorical feature DriveTrain which has three categories: *All*, *Front*, and *Rear*. Instead of the original training data, you are given the following table of summary statistics.

|  |  |  |  |
| --- | --- | --- | --- |
| **DriveTrain** | **Count** | **Mean** | **Corrected Sum of Squares** |
| All | 92 | 22.4673913043478 | 1574.9021739130400 |
| Front | 226 | 29.5044247787611 | 7794.4955752212400 |
| Rear | 110 | 25.0363636363636 | 983.8545454545450 |
| **Total** | **428** | **26.8434579439252** | **14074.5116822429000** |

# Question 8 (5 points)

You live in the San Francisco Bay area where earthquakes are not uncommon. Your house has a security alarm system against burglary, and it can be set off occasionally by an earthquake. Historically, there is a 6% chance that your house will be burglarized and there is a 2% chance that an earthquake will occur in your area. You can assume that the occurrences of burglary and earthquake are statistically independent. Based on your experience, your alarm will sound if the following events have occurred.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Earthquake** | True | True | False | False |
| **Burglary** | True | False | True | False |
| **Probability that Alarm will sound** | 0.99 | 0.15 | 0.95 | 0.0001 |

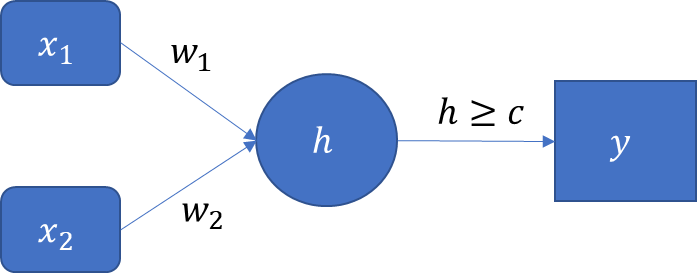
Please calculate this quantity Prob(Burglary = True and Earthquake = False | Alarm Sounded = False), i.e., the conditional probability that your house has been burglarized but no earthquake has occurred provided the alarm has sounded.

# Question 9 (5 points)

Logical operators (i.e., AND, OR, XAND, etc.) are the building blocks of any computational device. Logical functions return only two possible values, TRUE or FALSE, based on the truth or false values of their input values. For example, the operator OR returns FALSE only when all the input values are FALSE. Otherwise, the operator OR returns TRUE. If we denote TRUE by 1 and FALSE by 0, then the logical OR function can be represented by the following table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 0 | 0 | 1 | 1 |
|  | 0 | 1 | 0 | 1 |
| OR | 0 | 1 | 1 | 1 |

This function can be implemented by a perceptron with two binary inputs:



The activation functions for all the layers have this form: if . Otherwise, .

If we restrict the values of the parameters , , and to positive integers which are relatively prime (there is no integer greater than one that divides them), then specify these parameters such that the perceptron can implement the logical OR function.

# Question 10 (5 points)

The following table shows the observed target values and the predicted event probabilities from a model. The target is a binary variable whose values are Event and Non-Event. Please determine the threshold value that yields the highest F1 score. Please give your exact answer.

You should refrain from using the sklearn.metrics.precision\_recall\_curve function because it does not return all the possible thresholds.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Observed Target Value** | Non-Event | Non-Event | Event | Event | Event | Non-Event | Non-Event | Event | Event | Non-Event |
| **Predicted Event Probability** | 0.2 | 0.6 | 0.5 | 0.8 | 0.6 | 0.4 | 0.4 | 0.7 | 0.7 | 0.5 |

# Question 11 to 13

Please use the following information for answering Questions 11, 12, and 13.

We are interested in studying the effects of the Vehicle Age on the Claim Indicator. We particularly want to optimally separate the Vehicle Age into two groups.

You are given a two-way table and are asked to build a decision tree model using the Entropy criterion to discover the groups. We will treat the Vehicle Age as an **ordinal** predictor and the Claim Indicator as a **nominal** target variable. The order of the Vehicle Age is ‘1 to 3’ < ‘4 to 7’ < ‘8 to 10’ < ’11 and Above’.

|  |  |  |
| --- | --- | --- |
| **Vehicle Age (Number of Years)** | **Claim Indicator** | |
| No | Yes |
| 1 to 3 | 1,731 | 846 |
| 4 to 7 | 1,246 | 490 |
| 8 to 10 | 1,412 | 543 |
| 11 and Above | 2,700 | 690 |

# Question 11 (5 points)

What is the Entropy value of the root node?

# Question 12 (5 points)

Which is the most optimal way to form the two groups?

Multiple Choice:

1. {1 to 3} + {4 to 7, 8 to 10, 11 and Above}
2. {1 to 3, 4 to 7} + {8 to 10, 11 and Above}
3. {1 to 3, 4 to 7, 8 to 10} + {11 and Above}
4. {1 to 3, 8 to 10} + {4 to 7, 11 and Above}
5. {4 to 7, 8 to 10} + {1 to 3, 11 and Above}

# Question 13 (5 points)

What is the reduction in entropy of your most optimal split?

# Question 14 to 19

Please use the following information for answering Questions 14 to 19.

The Center for Machine Learning and Intelligent Systems at the University of California, Irvine manages the Machine Learning Repository (<https://archive.ics.uci.edu/ml/index.php>). You will train three models using the training CSV file WineQuality\_Train.csv. Then, you will select the model that yields the lowest misclassification rate on the test CSV file WineQuality\_Test.csv.

Our target variable is quality\_grp. It has two distinct values: 0 and 1.

Our input features are alcohol, citric\_acid, free\_sulfur\_dioxide, residual\_sugar, and sulphates. These five features are considered interval variables.

If the model algorithm can return predicted probabilities, then we will classify an observation to quality\_grp = 1 if the Prob(quality\_grp = 1) 0.1961733010776.

# Question 14 (5 points)

Build a Multinomial Logistic model with the following specifications.

* Include the Intercept term
* Use the Newton-Raphson optimization method
* Set the relative error in parameters acceptable for convergence to 0.00000001
* The algorithm must converge

What is the misclassification rate on the test data?

# Question 15 (5 points)

Build a Support Vector Machine classifier model with the following specifications.

* Specify the kernel to be linear
* Set the random state to 20191203
* Do not set a limit on the maximum number of iterations
* Disable probability estimates

What is the misclassification rate on the test data?

# Question 16 (5 points)

Build a Multi-layer Perceptron classifier model with the following specifications.

* Try the number of hidden layers from 1 to 10 by an increment of 1
* Try the number of hidden neurons per layer from 5 to 10 by an increment of 1
* Use the Rectified Linear Unit activation function for the hidden layers
* Use the lbfgs optimizing solver
* Specify the initial learning rate to 0.1
* Set the maximum number of iterations to 5000
* Set the random state value to 20191203

You will choose the configuration (i.e., number of layers and number of neurons) that yields the smallest loss on the training data. What is the misclassification rate on the test data?

# Question 17 (5 points)

Among the three models, multinomial logistic, Support Vector Machine classifier, and Multi-layer Perceptron classifier, which model gives the lowest misclassification rate on the test data?

Multiple Choice:

1. Multinomial Logistic
2. Support Vector Machine classifier
3. Multi-layer Perceptron classifier
4. Two-way Tie Multinomial Logistic and Multi-layer Perceptron classifier
5. Three-way Tie

# Question 18 (5 points)

Suppose you apply the Bagging technique on the multinomial logistic model with the specifications in Question 14. You tried the number of Bagging steps from 1 to 200 using your lucky random seed. After you have performed each number of Bagging steps, you calculate the misclassification rate on the test data. At the end of this experiment, you plot the misclassification rate on the test data against the number of Bagging steps. What will you expect of the graph?

Multiple Choice:

1. An upward trend going to 0.5
2. A downward trend going to 0.2
3. A trend converging to approximately 0.245
4. A trend converging to approximately 0.275
5. Inconclusive

# Question 19 (5 points)

You trained a classification tree on the Training data with the following specifications.

* The Splitting Criterion is the Entropy
* The maximum depth is 5
* The random state value is 20191203

Using the threshold of 0.5 on the predicted probability Prob(quality\_grp = 1), the classification tree yields an accuracy of 0.8159 on the Test data.

Next, you are going to apply the Adaptive Boosting technique on the above classification tree model with the following specifications.

* The maximum number of Boosting iterations is 50
* Terminate the iteration if the accuracy of the classification tree on the Training data is greater than or equal to 0.9999999
* If the observed quality\_grp is 1, then the absolute error is 1 – Prob(quality\_grp = 1). Otherwise, the absolute error is Prob(quality\_grp = 1).
* If an observation is correctly classified, then the weight is the absolute error. Otherwise, the weight is 2 plus the absolute error.

What is the accuracy of the boosted classification tree on the test data?

Multiple Choice:

1. 0.7862
2. 0.8328
3. 0.2072
4. 1.0000
5. 0.8159

# Question 20 (5 points)

If the Area Under Curve metric is one for a binary classification problem, then what can you conclude that the value of the Root Average Square Error metric?

Multiple Choice:

1. Equal to 0
2. Equal to 0.5
3. Greater than 0.5
4. Less than 0.5
5. Inconclusive