CS 584 Machine Learning

Autumn 2019 Midterm Test Answer Key

# Question 1 (5 points)

The expected confidence of an association rule X → Y is numerically equal to the support of the antecedent. True or False?

Answer: False. The expected confidence is equal to the support of the consequent.

# Question 2 (5 points)

If the Root Average Squared Error is zero in a binary classification problem, then the Area Under Curve must also be one. True or False?

Answer: True. When the Root Average Squared Error is zero, then the predicted event probability is 1 for when the observed target value is an event and the predicted event probability is 0 when the observed target value is a non-event. Therefore, there is one pair of probabilities and it is a concordant pair. Thus, the Area Under Curve must be one.

# Question 3 (5 points)

While you are building a multinomial logistic model for a label that has five categories, you encountered the complete separation phenomenon. Then which of the following may also happen? Select all those apply.

(A) The likelihood value of the model is practically one

(B) The root average squared error is practically zero

(C) All the standard errors are exceptionally huge

(D) Floating-point arithmetic overflow and/or underflow exceptions thrown

(E) The area under the curve is one

Answer: All except (E) because the area under the curve does not apply to a target variable that has more than two categories.

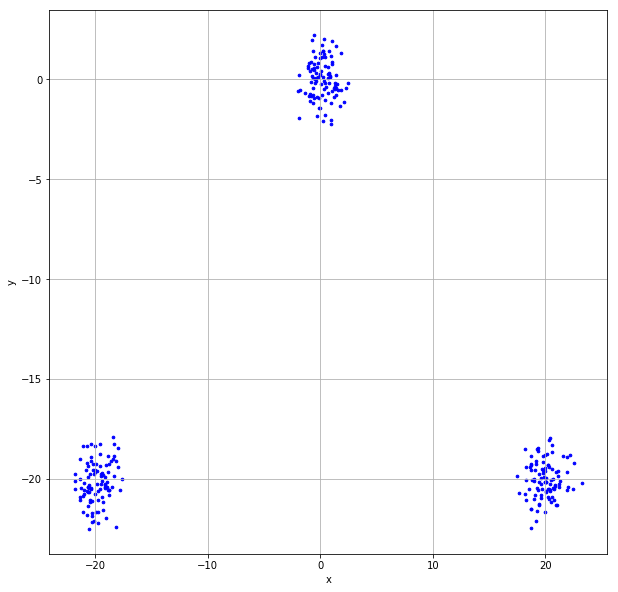
# Question 4 (5 points)

A nominal target variable has four categories. What is the theoretical highest entropy we may see in a node? (A) 0, (B) 1, (C) -2, (D) 0.75, (E) 2

Answer: (E). A node’s entropy will attain the theoretical maximum when the node is completely impure (i.e., the four categories have equal counts). In this case, the and the entropy is .

# Question 5 (5 points)

You have generated the following scatterplot of two fields x and y. Suppose you are going to perform the K-means clustering analysis on the data in the scatterplot. What do you think the Silhouette value will be for the 3-cluster solution?



(A) Close to negative one, (B) About zero, (C) Close to one, (D) Close to three, (E) Cannot be determined

Answer: (C) Close to one. It is because the scatterplot clearly shows 3 clusters of observation. Thus, the silhouette value for the 3-cluster solution should be close to one. Indeed, the silhouette value is 0.9368.

# Question 6 (5 points)

Suppose there are 100 unique items in the universal set, how many 5-itemset can you possibly generate?

Answer: The possible number of 5-itemset is

# Question 7 (5 points)

What is the Interquartile Range (IQR) value of a series of 101 consecutive integers: 1 to 101?

Answer: The first quartile is Q1 = 26 and the third quartile is Q3 = 76. Therefore, the interquartile range is IQR = Q3 – Q1 = 76 – 26 = 50.

# Question 8 (5 points)

Suppose we build a classification tree using a dataset with 1,000 observations. The target variable has five categories whose frequencies are listed below. What is the entropy value of the root node? Please round your answer to the seventh decimal place.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Target Category** | I | II | III | IV | V |
| **Frequency** | 64 | 250 | 364 | 259 | 63 |

Answer: The following table shows the observed proportions.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Target Category** | I | II | III | IV | V |
|  | 0.064 | 0.250 | 0.364 | 0.259 | 0.063 |
|  | -0.2538102 | -0.5 | -0.5307082 | -0.5047848 | -0.2512758 |

The entropy value of the root node is .

# Question 9 (5 points)

Suppose we build a multinomial logistic model using a dataset with 1,000 observations. The model contains only the Intercept terms. The target variable has five categories whose frequencies are listed below. The reference target category is Category III. Please estimate the Intercept of Category II. Please round your answer to the seventh decimal place.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Target Category** | I | II | III | IV | V |
| **Frequency** | 64 | 250 | 364 | 259 | 63 |

Answer: The estimate of the Intercept of Category V is the natural logarithm of the ratio of the count for Category II to the count for Category III (the reference). Therefore, the Intercept of Category II is .

# Question 10 (5 points)

You invited your six friends to your home to watch a basketball game. Your friends brought snacks and beverages along. The following table lists the items your friends brought.

|  |  |
| --- | --- |
| Friend | Items |
| Andrew | Cheese, Cracker, Soda, Wings |
| Betty | Cheese, Soda, Tortilla |
| Carl | Cheese, Ice Cream, Soda, Wings |
| Danny | Cheese, Ice Cream, Salsa, Tortilla |
| Emily | Salsa, Soda, Tortilla, Wings |
| Frank | Cheese, Cracker, Ice Cream, Soda, Wings |

You noticed that many of your friends brought Cheese, Soda, and Wings together. Since you rather want to spend your money on food than Soda, you want to study how likely your friends will also bring Soda if they are going to bring Cheese and Wings. Therefore, please determine the Lift of this association rule {Cheese, Wings} ==> {Soda}. Please round your answer to the seventh decimal place.

Answer: The Conference of this rule is Pr(Cheese, Soda, Wings) / Pr(Cheese, Wings) = (3/6) / (3/6) = 1. The Expected Conference of this rule is Pr(Soda) = 5/6. Therefore, the Lift = Conference / (Expected Conference) = 1 / (5/6) = 6/5 = 1.2.

# Question 11 (5 points)

Given the following confusion matrix

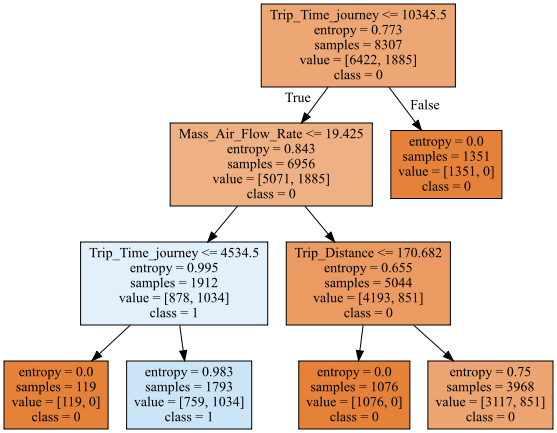
|  |  |  |
| --- | --- | --- |
|  | Predicted Event | Predicted Non-Event |
| Observed Event | 18 | 7 |
| Observed Non-Event | 12 | 22 |

What are the Precision and the Recall values? Please give your answers as a fraction between 0 and 1 inclusively.

Answer: Precision = 18 / (18 + 12) = 0.6 and Recall = 18 / (18 + 7) = 0.72.

# Question 12 (10 points)

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



Suppose we are interested in predicting the class 1 of the target variable. In other words, the predicted event probability is the Prob (class = 1). What is the Area Under Curve value? Please round your answer to the seventh decimal place.

Answer: 0.8046672.

The decision tree has five leaves. The following table summarizes the number of observations and the predicted event probability.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Leaf | 1 | 2 | 3 | 4 | 5 |
| Number of Non-Event Observations | 1351 | 119 | 759 | 1076 | 3117 |
| Number of Event Observations | 0 | 0 | 1034 | 0 | 851 |
| Predicted Event Probability | 0 | 0 | 1034/(1034+759)  =0.5767 | 0 | 851/(851+3117)  =0.2145 |

Based on the above table, we can generate the following table.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | |  | Non-Event Observation | | |
| N = 2546 =  119 + 1076+1351 | N = 3117 | N = 759 |
| Predicted Event Probability | 0 | 0.2145 | 0.5767 |
| Event Observation | N = 851 | 0.2145 | Concordant | Tied | Discordant |
| N = 1034 | 0.5767 | Concordant | Concordant | Tied |

The number of Concordant pairs is 851\*2546 + 1034\*2546 + 1034\*3117 = 8,022,188. The number of Discordant pairs is 851\*759 = 645,909. The number of Tied pairs is 851\*3117 + 1034\*759 = 3,437,373. Therefore, the Area Under Curve value is 0.5 + 0.5 \* (8,022,188 - 645,909) / (8,022,188 + 645,909 + 3,437,373) = 0.5 + 0.5 \* 7376279 / 12105470 = 0.8046672.

# Question 13 (10 points)

You are going to build a logistic model using the 20 observations below. The binary target variable is y, and the interval predictor is x.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| x | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 | 4 | 4 | 4 | 4 |
| y | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |

The specifications are:

1. The Intercept term is included
2. The optimization method is Newton
3. The maximum number of iterations is 100
4. The tolerance level is 1e-8.

After you have built your model, you will apply them on the following test data and then calculate the root average squared error metric.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| x | 0 | 1 | 2 | 3 | 4 |
| y | 0 | 0 | 1 | 1 | 1 |

What is the Root Average Squared Error when the Logistic model is applied to the test data? Please round your answer to the seventh decimal place.

Answer: For the test data, the root average squared error of the Logistic Model is 0.2493392.

# Question 14 (10 points)

You can use Chicago’s 311 Service Request to report street potholes. After a request has been received, the Department of Transportation will first assess the severity of the pothole, and then schedule a road crew to fill up the pothole. After the pothole is filled, the service request will be closed.

You are provided with this CSV file **ChicagoCompletedPotHole.csv** for analyzing the city’s efforts to fill up street potholes. The data contains 17,912 observations. Each observation represents a completed request which was created between December 1, 2017 and March 31, 2018 and was completed between December 4, 2017 and September 12, 2018. The data has the following seven variables:

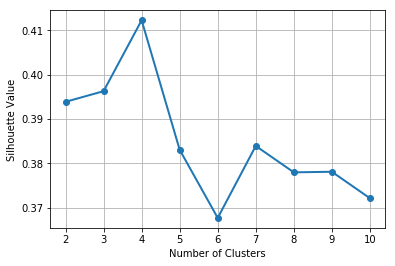
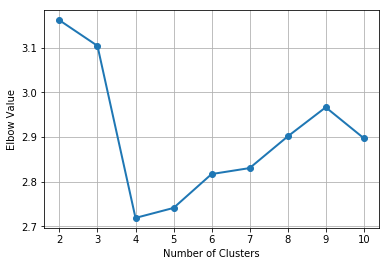
|  |  |  |
| --- | --- | --- |
| **Name** | **Level** | **Description** |
| 1. CASE\_SEQUENCE | Nominal | A unique index for identifying an observation |
| 1. WARD | Nominal | Chicago’s ward number from 1 to 50 |
| 1. CREATION\_MONTH | Nominal | Calendar month when the request was created |
| 1. N\_POTHOLES\_FILLED\_ON\_BLOCK | Interval | Number of potholes filled on the city block |
| 1. N\_DAYS\_FOR\_COMPLETION | Interval | Number of days elapsed until completion |
| 1. LATITUDE | Interval | Latitude of the city block |
| 1. LONGITUDE | Interval | Longitude of the city block |

You will use the K-Means Clustering algorithm to identify clusters in the entire data with the following specifications.

1. Use loge(N\_POTHOLES\_FILLED\_ON\_BLOCK), loge(1 + N\_DAYS\_FOR\_COMPLETION), LATITUDE, and LONGITUDE (i.e., you need to perform the transformations before clustering)
2. The maximum number of clusters is 10 and the minimum number of clusters is 2
3. The random seed is 20191009
4. Use both the Elbow and the Silhouette methods to determine the number of clusters

What is the number of clusters? What are the Elbow and the Silhouette values? Except for the number of clusters, please round your answers to SEVEN decimal places.

Answer: The Elbow chart shows an elbow at the 4-cluster solution and the Silhouette chart shows a peak at the 4-cluster solution.



Therefore, we determine the number of clusters as 4. The Elbow value is 2.7189805. The Silhouette value is 0.4122889.

# Question 15 and Question 16 (15 points)

In the automobile industry, a common question is how likely a policy-holder will file a claim during the coverage period. Your task is to build a logistic model. To avoid discriminating policy-holders, we will use predictors that can be verified and are related to the risk exposures of the policy-holders. The CSV file policy\_2001.csv contains data about 617 policy-holders. We will use only the following variables.

**Target Variable**

* + CLAIM\_FLAG: Claim Indicator (1 = Claim Filed, 0 = Otherwise) and 1 is the event value.

**Nominal Predictor**

* + CREDIT\_SCORE\_BAND: Credit Score Tier (‘450 – 619’, ‘620 – 659’, ‘660 – 749’, and ‘750 +’)

**Interval Predictors**

* + BLUEBOOK\_1000: Blue Book Value in Thousands of Dollars (min. = 1.5, max. = 39.54)
  + CUST\_LOYALTY: Number of Years with Company Before Policy Date (min. = 0, max. ≈ 21)
  + MVR\_PTS: Motor Vehicle Record Points (min. = 0, max. = 10)
  + TIF: Time-in-Force (min. = 101, max. = 107)
  + TRAVTIME: Number of Miles Distance Commute to Work (min. = 5, max. ≈ 93)

Since the tools may not take the nominal predictor as is, you will first derive the dummy indicators from the nominal predictors and then use the dummy indicators in building the models. You will build the logistic model according to the following specifications.

* + The optimization algorithm is the Newton-Raphson method
  + The maximum number of iterations is 100
  + The relative error in parameter estimates acceptable for convergence is 1E-8
  + The Intercept term must be included in the model
  + Use the Forward Selection method to enter predictors, the alpha level is 0.05.

You will divide the data into the Training and the Test partitions. You will build the logistic model using the Training partition. Later, you will evaluate the model based on the Test partition.

**Data Partition**

* + The Training partition consists of 75% of the original observations, the remaining 25% goes to the Test partition.
  + Use the CLAIM\_FLAG as the stratum variable.
  + The random seed is 20191009.

# Question 15 (10 points)

What predictors are selected for the logistic model?

Answer: The following table summarizes the Forward Selection process.

| Step | Model | Number of Free Parameter | Log-Likelihood | Deviance | Degrees of Freedom | Significance |
| --- | --- | --- | --- | --- | --- | --- |
| 0 | Intercept | 1 | -277.3115456 |  |  |  |
| 1.1 | Intercept + CREDIT\_SCORE\_BAND | 4 | -277.1179381 | 0.3872150 | 3 | 0.9428707 |
| 1.2 | Intercept + BLUEBOOK\_1000 | 2 | -275.7574537 | 3.1081839 | 1 | 0.0778998 |
| 1.3 | Intercept + CUST\_LOYALTY | 2 | -276.8175928 | 0.9879055 | 1 | 0.3202548 |
| 1.4 | Intercept + MVR\_PTS | 2 | -264.2842244 | 26.0546424 | 1 | 0.0000003 |
| 1.5 | Intercept + TIF | 2 | -277.2029003 | 0.2172907 | 1 | 0.6411121 |
| 1.6 | Intercept + TRAVTIME | 2 | -275.1286268 | 4.3658377 | 1 | 0.0366665 |
|  |  |  |  |  |  |  |
| 1 | Intercept + MVR\_PTS | 2 | -264.2842244 |  |  |  |
| 2.1 | Intercept + MVR\_PTS + CREDIT\_SCORE\_BAND | 5 | -263.7347730 | 1.0989029 | 3 | 0.7773389 |
| 2.2 | Intercept + MVR\_PTS + BLUEBOOK\_1000 | 3 | -263.1946085 | 2.1792319 | 1 | 0.1398843 |
| 2.3 | Intercept + MVR\_PTS + CUST\_LOYALTY | 3 | -264.0018538 | 0.5647412 | 1 | 0.4523562 |
| 2.4 | Intercept + MVR\_PTS + TIF | 3 | -264.2692459 | 0.0299571 | 1 | 0.8625877 |
| 2.5 | Intercept + MVR\_PTS + TRAVTIME | 3 | -262.0566029 | 4.4552431 | 1 | 0.0347943 |
|  |  |  |  |  |  |  |
| 2 | Intercept + MVR\_PTS + TRAVTIME | 3 | -262.0566029 |  |  |  |
| 3.1 | Intercept + MVR\_PTS + TRAVTIME + CREDIT\_SCORE\_BAND | 6 | -261.5556277 | 1.0019504 | 3 | 0.8007800 |
| 3.2 | Intercept + MVR\_PTS + TRAVTIME + BLUEBOOK\_1000 | 4 | -260.8403356 | 2.4325346 | 1 | 0.1188406 |
| 3.3 | Intercept + MVR\_PTS + TRAVTIME + CUST\_LOYALTY | 4 | -261.8310196 | 0.4511666 | 1 | 0.5017815 |
| 3.4 | Intercept + MVR\_PTS + TRAVTIME + TIF | 4 | -262.0138588 | 0.0854882 | 1 | 0.7699934 |

Since the Significances of all the predictors in Step 3 are greater than 0.05, no additional predictors are selected, and the Forward Selection process stops at Step 3.

The final model is Intercept + MVR\_PTS + TRAVTIME.

# Question 16 (5 points)

Using the claim rate in the Training partition as the threshold for predicted claim probability, calculate the misclassification rate of the logistic model on the Test partition? Please round your answers to SEVEN decimal places.

Answer: The threshold for predicted claim probability is 0.2878788. The Confusion Matrix based on the Test partition is

|  |  |  |
| --- | --- | --- |
| Observed CLAIM\_FLAG | Predicted CLAIM\_FLAG | |
| 0 | 1 |
| 0 | 80 | 31 |
| 1 | 27 | 17 |

The Misclassification Rate is (31 + 27) / 155 = 0.3741935.