CS 584 Machine Learning

Spring 2019 Midterm Test Answer Key

# Question 1 (5 points)

What is the Interquartile Range (IQR) value of this series of eleven values: 0.1811, 0.0775, 0.1279, 0.0045, 0.0001, 0.9457, 0.0021, 0, 0.0005, 0.7305, and 0.8936? Please give your answer up to four decimal places.

Answer: You may be surprised if I tell you that there are multiple answers for the first and the third quartile, in particular when the sample size is small. In our case, the sample size is 11. Since 11\*0.25 = 2.75 and 11\*0.75 = 8.25, the first quartile is the 2.75-th observation and the third quartile is the 8.25-th observation. The ambiguity is how to define where the fractionally-indexed observations are.

If you accept that the first quartile is the median of the lower half data and the third quartile is the median of the upper half data, then the first quartile Q1 = 0.0005 and the third quartile Q3 = 0.7305. Therefore, the interquartile range IQR = 0.7305 – 0.0005 = 0.73.

If you use the Percentile function in the Numpy package, then the first quartile Q1 = 0.0013 and the third quartile Q3 = 0.4558. Therefore, the interquartile range IQR = Q3 – Q1 = 0.4558 – 0.0013 = 0.4545.

# Question 2 (5 points)

Suppose we compute the Pearson Chi-Squared statistic to determine if the *Claim Indicator* is statistical independent of the *Number of Children Driving*. What is the Pearson Chi-Squared statistic value and the corresponding one-sided significance (i.e., the p-value)? Please give your answer in scientific notation in this format n.nnnnE-mm.

|  |  |  |
| --- | --- | --- |
| **Number of Children Driving** | **Claim Indicator** | |
| No | Yes |
| 0 | 6,815 | 2,254 |
| 1 | 492 | 312 |
| 2 | 212 | 139 |
| 3+ | 37 | 41 |

Answer: Let , and . For example, . Let the column marginal counts be , . Let the row marginal counts be , . Finally, the total counts be Let , and be the expected count under the independence assumption. The following table shows the expected counts.

|  |  |  |
| --- | --- | --- |
| **Number of Children Driving** | **Claim Indicator** | |
| No | Yes |
| 0 | 6651.6564 | 2417.3436 |
| 1 | 589.6937 | 214.3063 |
| 2 | 257.4409 | 93.5591 |
| 3+ | 57.2091 | 20.7909 |

The Pearson Chi-Squared statistic is = 132.6414 with degrees of freedom. The significance value is = 1.4582e-28.

# Question 3 (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 give your answer up to four decimal places.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **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 |
|  | -0.2538 | -0.5 | -0.5307 | -0.5048 | -0.2513 |

The entropy value of the root node is .

# Question 4 (5 points)

For a nominal target variable with five categories, what is the maximum Gini’s value in theory? Please give your answer up to four decimal places.

Answer: The maximum Gini’s value is attained when the probabilities are equal. In other words, each probability is 1/5. Therefore, the maximum Gini’s value = .

# Question 5 (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 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 = True), i.e., the conditional probability that your house has been burglarized but no earthquake has occurred provided the alarm has sounded. Please provide your answer up to four decimal places.

Answer: The following table shows all the joint probabilities.

| **Alarm** | **Burglary** | **Earthquake** | **Prob(Alarm | Burglary, Earthquake)** | **Prob(Burglary)** | **Prob(Earthquake)** | **Prob(Alarm, Burglary, Earthquake)** |
| --- | --- | --- | --- | --- | --- | --- |
| TRUE | TRUE | TRUE | 0.99 | 0.06 | 0.02 | **0.001188** |
| TRUE | FALSE | TRUE | 0.15 | 0.94 | 0.02 | **0.00282** |
| TRUE | TRUE | FALSE | 0.95 | 0.06 | 0.98 | **0.05586** |
| TRUE | FALSE | FALSE | 0.0001 | 0.94 | 0.98 | **0.00009212** |

The marginal probabilities for Alarm are

* Prob(Alarm = TRUE) = 0.001188 + 0.00282 + 0.05586 + 0.00009212 = 0.05996012
* Prob(Alarm = FALSE) = 1 – Prob(Alarm = TRUE) = 1 - 0.05996012 = 0.94003988.

The conditional probability Prob(Burglary = True and Earthquake = False | Alarm Sounded = True) = Prob(Burglary = True and Earthquake = False and Alarm Sounded = True) / Prob(Alarm = TRUE) = 0.05586 / 0.05996012 = 0.9316.

# Question 6 (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 V. Please give your answer up to four decimal places.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **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 V to the count for Category III (the reference). Therefore, the Intercept of Category V is .

# Question 7 (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 calculate the Area Under Curve statistics for this model. Please give your answer up to four decimal places.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Observed Target Value** | Event | Non-Event | Non-Event | Event | Event | Non-Event | Event | Non-Event | Event | Event |
| **Predicted Event Probability** | 0.8 | 0.5 | 0.4 | 0.6 | 0.4 | 0.7 | 0.0 | 0.5 | 0.7 | 0.6 |

Answer: The following table shows the 24 pairs of predicted event probabilities. The rows are of the observed Event and the columns are of the observed Non-Event. C denotes a concordant pair, D a discordant pair, and T a tied pair.

| **Observed Event** | **Observed Non-Event** | | | |
| --- | --- | --- | --- | --- |
| **0.4** | **0.5** | **0.5** | **0.7** |
| **0** | D | D | D | D |
| **0.4** | T | D | D | D |
| **0.6** | **C** | **C** | **C** | D |
| **0.6** | **C** | **C** | **C** | D |
| **0.7** | **C** | **C** | **C** | T |
| **0.8** | **C** | **C** | **C** | **C** |

There are 13 concordant pairs, 9 discordant pairs, and 2 tied pairs. The Area Under Curve statistic is equal to 0.5 + 0.5 \* (13 – 9) / 24 = 0.5 + 0.5 \* 4 / 24 = 0.5833.

# Question 8 (5 points)

Using the table shown in Question 7, please calculate the Misclassification Rates for that model using the observed Event proportion as the threshold. If the predicted event probability is greater than or equal to the threshold, then an event is predicted, otherwise a non-event is predicted. Please give your answer up to four decimal places.

Answer: There are 6 observed Event out of the 10 observations. Therefore, the observed Event proportion is 0.6. The following table contains the results of the classification.

| **Observed Target Value** | Event | Non-Event | Non-Event | Event | Event | Non-Event | Event | Non-Event | Event | Event |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Predicted Event Probability** | 0.8 | 0.5 | 0.4 | 0.6 | 0.4 | 0.7 | 0.0 | 0.5 | 0.7 | 0.6 |
| **Predicted Target Value** | Event | Non-Event | Non-Event | Event | Non-Event | Event | Non-Event | Non-Event | Event | Event |
| **Correctly Classified** | Yes | Yes | No | Yes | Yes | No | No | Yes | Yes | Yes |

Since three out of ten observations are misclassified, the Misclassification Rate = 3 / 10 = 0.3.

# Question 9 (5 points)

Suppose there are 50 unique items in the universal set, how many 4-itemset can you possibly generate?

Answer: The possible number of 4-itemset is

# Question 10 (5 points)

We built a model to predict the outcome of a binary target variable. Based on the predicted event probabilities for 4,217 observations, we have partially completed the Gain and Lift Summary table. The Decile N and the Gain N rows are filled in as below. Please calculate the Accumulated Lift value of the fourth decile (i.e., Decile 4). Please give your answer up to four decimal places.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Decile** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| **Decile N** | 421 | 422 | 422 | 422 | 421 | 422 | 422 | 422 | 422 | 421 |
| **Gain N** | 155 | 52 | 26 | 19 | 22 | 27 | 24 | 19 | 18 | 22 |

Answer: The accumulated Gain and Lift Summary table is as below.

| **Decile** | **Acc. Decile N** | **Acc. Decile %** | **Acc. Gain N** | **Acc. Gain %** | **Acc. Response %** | **Acc. Lift** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | 421 | 10% | 155 | 40.4% | 36.8% | 4.0432 |
| 2 | 843 | 20% | 207 | 53.9% | 24.6% | 2.6966 |
| 3 | 1265 | 30% | 233 | 60.7% | 18.4% | 2.0227 |
| 4 | 1687 | 40% | 252 | 65.6% | 14.9% | 1.6404 |
| 5 | 2108 | 50% | 274 | 71.4% | 13.0% | 1.4274 |
| 6 | 2530 | 60% | 301 | 78.4% | 11.9% | 1.3065 |
| 7 | 2952 | 70% | 325 | 84.6% | 11.0% | 1.2090 |
| 8 | 3374 | 80% | 344 | 89.6% | 10.2% | 1.1197 |
| 9 | 3796 | 90% | 362 | 94.3% | 9.5% | 1.0473 |
| 10 | 4217 | 100% | 384 | 100.0% | 9.1% | 1.0000 |

The accumulated lift value of the fourth decile is 1.6404.

# Question 11 (25 points)

You can use the 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 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 first identify clusters in the data, and then use a classification tree to profile the clusters. Here are the specifications for performing the analyses.

**K-Means Clustering**

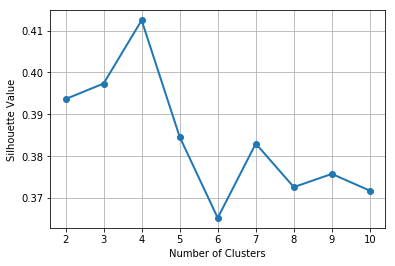
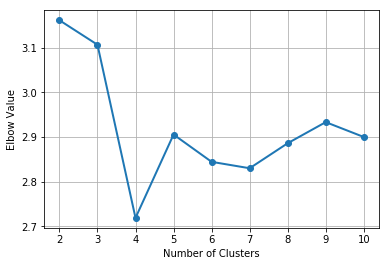
1. Use loge(N\_POTHOLES\_FILLED\_ON\_BLOCK), loge(1 + N\_DAYS\_FOR\_COMPLETION), LATITUDE, and LONGITUDE (i.e., you need to perform the transformations before clustering)
2. The maximum number of clusters is 10 and the minimum number of clusters is 2
3. The random seed is 20190327
4. Use both the Elbow and the Silhouette methods to determine the number of clusters

**Classification Tree**

1. Use N\_POTHOLES\_FILLED\_ON\_BLOCK, N\_DAYS\_FOR\_COMPLETION, LATITUDE, and LONGITUDE (without any transformations) as the predictors
2. The maximum number of branches is 2
3. The maximum depth is 2
4. The random seed is 20190327.
5. The grow criterion is the Gini’s value

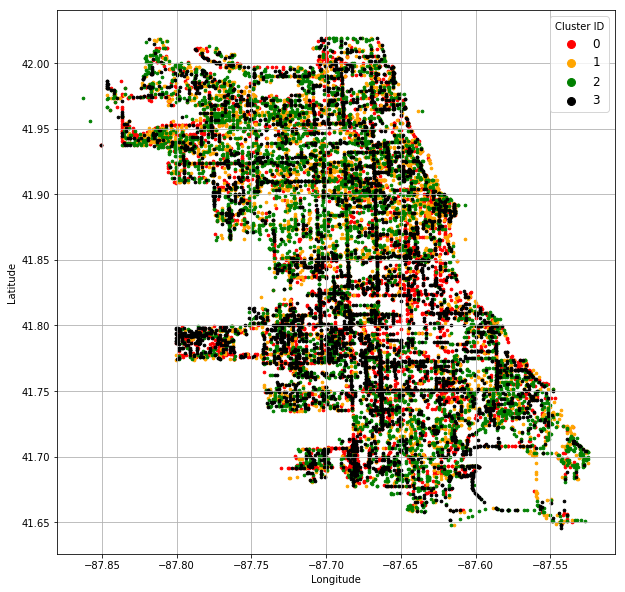
Please answer the following questions.

1. (5 points) How many clusters have you determined? Please provide the Elbow and the Silhouette charts and state your arguments. The charts must be properly labeled.



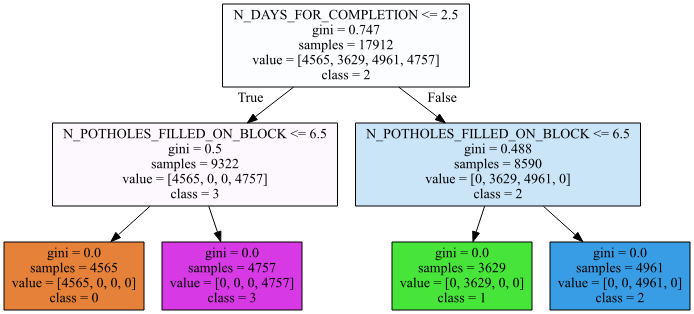
The Elbow chart shows an elbow at the 4-cluster solution and the Silhouette chart shows a peak at the 4-cluster solution. Therefore, we determine the number of clusters as 4.

1. (5 points) Generate a scatterplot of LATITUDE (y-axis) versus LONGITUDE (x-axis) using the Cluster ID as the color response variable. You may need to adjust the marker size and set the aspect ratio to one in order to make the scatterplot more readable.



1. (5 points) How many leaves in your classification tree? Please provide the properly labeled tree diagram.

The classification tree has four leaves.



1. (5 points) What is the Root Average Squared Error of your classification tree? Please give your answer up to four decimal places.

Since the Gini value of each leaf is zero, this indicates that all observations in each leaf come from the same cluster and they are completely and correctly classified. Therefore, the Root Average Squared Error is zero.

1. (5 points) Based on your classification tree, please describe the profiles of the clusters which are at least 99% correctly classified by the classification tree.

| **Cluster ID** | **Size** | **N\_DAYS\_FOR\_COMPLETION** | **N\_POTHOLES\_FILLED\_ON\_BLOCK** |
| --- | --- | --- | --- |
| 0 | 4,565 | <= 2.5 | <= 6.5 |
| 1 | 3,629 | > 2.5 | <= 6.5 |
| 2 | 4,961 | > 2.5 | > 6.5 |
| 3 | 4,757 | <= 2.5 | > 6.5 |

* Cluster 0 has 4,565 reports. In those reports, 6 or less potholes were filled on the block within 2 days of receiving the reports.
* Cluster 1 has 3,629 reports. In those reports, 6 or less potholes were filled on the block but after at least 3 days of receiving the reports.
* Cluster 2 has 4,961 reports. In those reports, 7 or more potholes were filled on the block after at least 3 days of receiving the reports.
* Cluster 3 has 4,757 reports. In those reports, 7 or more potholes were filled on the block within 2 days of receiving the reports.

# Question 12 (25 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 two models. After evaluating and comparing the models, you will recommend the model that performs better. In order 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 two models according to the following specifications.

**Classification Tree Model**

* + The maximum number of depths is 5
  + The splitting criterion is Entropy
  + The random seed is 20190402

**Logistic Model**

* + 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

You will divide the data into the Training and the Testing partitions. You will build and evaluate the three models using the Training partition. Later, you will recommend one model based on the evaluation and the comparison results from the Testing partition.

**Data Partition**

* + The Training partition consists of 75% of the original observations, the remaining 25% goes to the Testing partition.
  + The claim rates (i.e., the fraction of observations whose CLAIM\_FLAG is 1) must be the same in both partitions.
  + The random seed is 20190402.

Please answer the following questions.

1. (5 points) How many observations are in the Training and the Testing partitions?

The Training partition has 462 observations and the Testing partition has 155 observations.

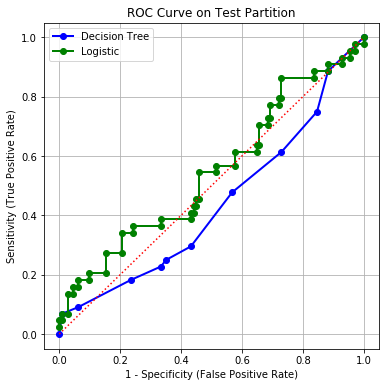
1. (5 points) What is the claim rate in the Training partition?

The claim rate in the Training partition is 133 / (329 + 133) = 0.2879.

1. (5 points) Use **the claim rate in the Training partition** as the probability threshold in the misclassification rate calculation. A claim is predicted if the predicted probability of filing a claim is greater than or equal to the probability threshold. Calculate the Area Under Curve metric, the Root Average Squared Error metric, and the Misclassification Rate for both models using the Testing partition. Present your results in a table, list the metrics in the column dimension and the models in the row dimension.

|  |  |  |
| --- | --- | --- |
|  | Classification Tree | Logistic |
| Area Under Curve | 0.4351 | **0.5463** |
| Root Average Squared Error | 0.5089 | **0.4539** |
| Misclassification Rate | **0.4581** | 0.4839 |

1. (5 points) Calculate (but no need to display) the coordinates of the Receiver Operating Characteristic curve for both models using the Testing partition. Plot both ROC curves in the same chart but uses a different color for each curve. The chart (including the axes, the title, and the curve legends) must be properly labeled.



1. (5 points) Based on the evaluation and the comparison results in (c) and (d), recommend your model. You must state your reasons for your recommendation.

The Logistic Regression model has the higher Area Under Curve metric and the lower Root Average Squared Error. On the other hand, the Decision Tree model has the lower Misclassification Rate. In the Receiver Operating Characteristic chart, the curve of the Logistic Regression model is above that of the Decision Tree model most, if not all, of the time. In fact, the curve of the Decision Tree model falls below the diagonal reference line the majority of time. This indicates that the Decision Tree is worse than a uniformly random model.

In summary, the Logistic Regression model wins 3 out of 4 criteria, and the Decision Tree model only win in the Misclassification Rate criterion. Therefore, I will recommend the Logistic Regression model.