CS 584-04: Machine Learning

Fall 2018 Midterm Test Answer Key

# Question 1 (50 points)

Anyone 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 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 | The 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 between CREATION\_DATE and COMPLETION\_DATE inclusively |
| 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 respective analyses.

**K-Means Clustering**

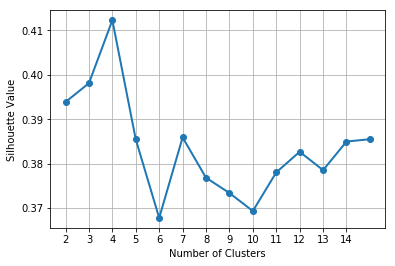
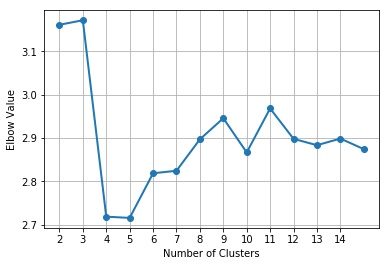
1. Use loge(N\_POTHOLES\_FILLED\_ON\_BLOCK), loge(1 + N\_DAYS\_FOR\_COMPLETION), LATITUDE, and LONGITUDE (i.e., you need to perform the transformations before clustering)
2. The maximum number of clusters is 15 and the minimum number of clusters is 2
3. The random seed is 20181010
4. Use both the Elbow and the Silhouette methods to determine the number of clusters

**Classification Tree**

1. The target variable is the Cluster ID
2. Use N\_POTHOLES\_FILLED\_ON\_BLOCK, N\_DAYS\_FOR\_COMPLETION, LATITUDE, and LONGITUDE (without any transformations) as the predictors
3. The maximum number of branches is 2
4. The maximum depth is 2
5. The random seed is 20181010.
6. The grow criterion is Entropy

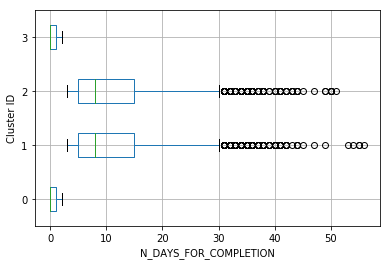
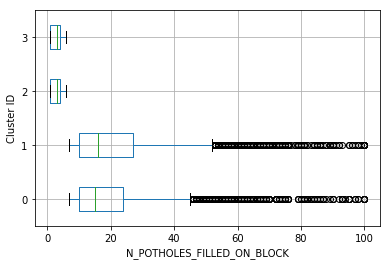
Please answer the following questions.

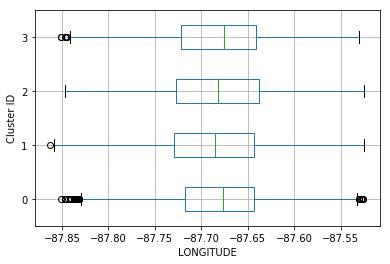
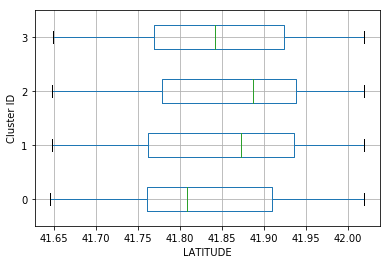
1. (10 points) How many clusters did you determine? Please provide the Elbow and the Silhouette charts and state your arguments. The charts must be properly labeled.



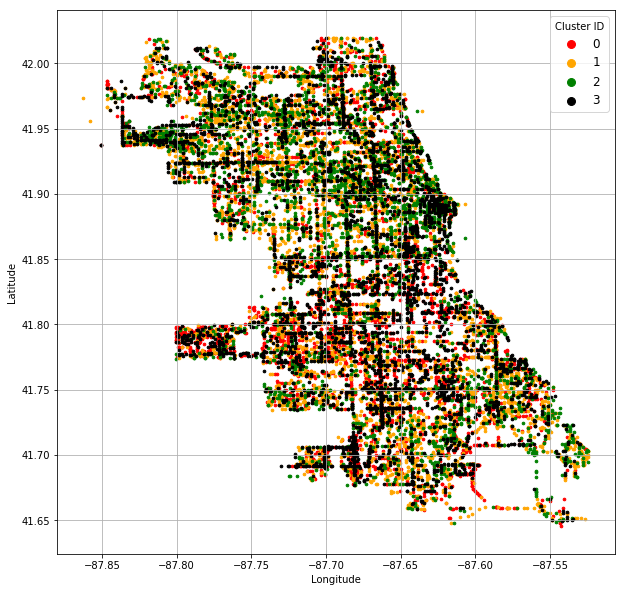
The Elbow chart shows an elbow at the 4-cluster solution and the Silhouette chart shows a peak at the 4-cluster solution. Therefore, we determine the number of clusters as 4.

1. (5 points) Create a box-plot for each of these four variables: N\_POTHOLES\_FILLED\_ON\_BLOCK, N\_DAYS\_FOR\_COMPLETION, LATITUDE, and LONGITUDE, grouped by the Cluster ID.





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.

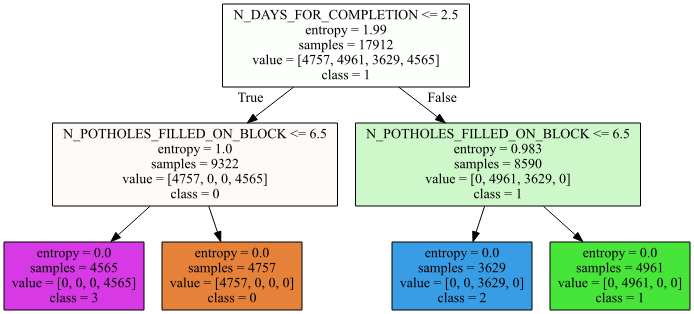


1. (5 points) Comment your scatterplot in (c). In particular, how effective or ineffective do you think the clustering analysis in dividing up the observations according to their geographical locations?

The scatterplot resembles a map of the city of Chicago. The clustering analysis is not effective in dividing up the observations according to their geographical locations. It is because the colors are all over the map. However, there are some regions which have more markers of a particular color. For example, the south and the southwest regions of Chicago has more Cluster 0 (red markers) than other regions. The Cluster 3 (black markers) concentrates at the downtown area and at the major roadways (horizontal and vertical black lines).

1. (10 points) How many leaves did you have in your classification tree? Please attach the tree diagram which must be properly labeled.

The classification tree has 4 leaves. Each leaf has zero entropy which indicates the Cluster IDs are constant within the leaf.



1. (5 points) Calculate the Misclassification Rate and the Root Average Squared Error of your classification tree.

Since the entropy of each leaf is zero, this indicates that each leaf contains the observations in a particular cluster. It follows that the Misclassification Rate is zero and the Root Average Squared Error is also zero.

1. (10 points) Based on your classification tree, how would you describe the profiles of the clusters?

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

# Question 2 (50 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 several 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 three models according to the following specifications.

**Nearest Neighbors Model**

* + The number of neighbors is 3
  + The distance metric is the standard Euclidean distance
  + The search algorithm is the brute-force method

**Classification Tree Model**

* + The maximum number of depths is 10
  + The splitting criterion is Entropy
  + The random seed is 20181010

**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 70% of the original observations, the remaining 30% 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 20181010

Please answer the following questions.

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

The Training partition has 431 observations and the Testing partition has 186 observations.

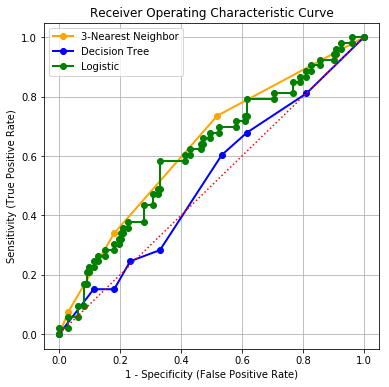
1. (5 points) What are the claim rates in the Training and the Testing partitions?

The claim rate in the Training partition is 124 / (307 + 124) = 0.287703. The claim rate in the Testing partition is 53 / (133 + 53) = 0.284946. Although we ideally expect both claim rates to be the same, they are different indeed due to rounding of the number of observations in creating the partitions.

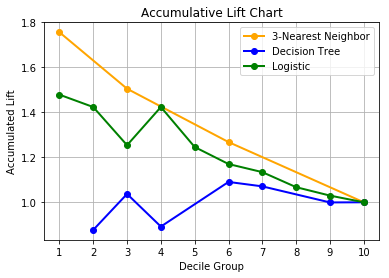
1. (10 points) Use the **claim rate in the Training partition** as the probability threshold in the misclassification rate calculation. A claim is predicted if the predicted probability of filing a claim is greater than or equal to the probability threshold. Calculate the Area Under Curve metric, the Root Average Squared Error metric, and the Misclassification Rate for all three models using the Testing partition. List the metrics as the rows and the models as the columns in a table.

|  |  |  |  |
| --- | --- | --- | --- |
|  | 3-Nearest Neighbors | Classification Tree | Logistic |
| Area Under Curve | **0.631934** | 0.513974 | 0.609590 |
| Root Average Squared Error | 0.479571 | 0.533012 | **0.446161** |
| Misclassification Rate | 0.446237 | 0.381720 | **0.360215** |

1. (10 points) Calculate (but no need to display) the coordinates of the Receiver Operating Characteristic curve for each of the three models using the Testing partition. Plot all three curves in the same chart but use a different color for each curve. The chart (including the axes, the title, and the curve legends) must be properly labeled.



1. (10 points) Calculate (but no need to display) the coordinates of the Accumulated Lift chart for each of the three models using the Testing partition. Plot all three accumulated lift curves in the same chart but use a different color for each curve. The chart (including the axes, the title, and the curve legends) must be properly labeled.



1. (10 points) Based on the evaluation and the comparison results in (c), (d), and (e), which single model will you recommend? Please state your reasons for your recommendation.

The 3-Nearest Neighbor model has the highest Area Under Curve metric. The Logistic model has the lowest Root Average Squared Error and the lowest Misclassification Rate. In the Receiver Operating Characteristic chart, the curve of the 3-Nearest Neighbor model is above that of the Logistic model most of the time. This finding is observed again in the Accumulative Lift chart. In summary, the 3-Nearest Neighbor model wins 3 out of 5 criteria, the Logistic model wins 2 out of 5 criteria, and the Decision Tree model loses in all 5 criteria. Therefore, I will recommend the 3-Nearest Neighbor model.

However, if I also consider the ease of deployment, then I will opt for my second choice which is the Logistic model. It is because any Nearest Neighbor model requires the entire training dataset to be accessible in deployment. Thus, this makes it a less desirable model.