University of Ottawa

CSI4142 Introduction to Data Science

Assignment 2 2018

Total: 70 marks

Instructions:
Submit your assignment thought the Virtual Campus. This is an individual assignment.

The goal of this assignment is to explore relevant concepts related to machine learning and data mining.

Answer all the following questions.

For questions 1 and 2, consider the following table that contains the partial data and schema of a Customer profiling database owned by a Bank. The Mortgage is the target class (or so-called label).

age	gender	income	children	mortgage
48	FEMALE	$17,546.00	1	NO
40	MALE	$30,085.10	3	YES
51	FEMALE	$16,575.40	0	NO
23	FEMALE	$20,375.40	3	NO
57	FEMALE	$50,576.30	0	NO
57	FEMALE	$37,869.60	2	NO
22	MALE	$8,877.07	0	NO
58	MALE	$24,946.60	0	NO
37	FEMALE	$2,500,304.30	2	NO
54	MALE	$24,212.10	2	NO
66	FEMALE	$59,803.90	0	NO
52	FEMALE	$26,658.80	0	YES
44	FEMALE	$15,735.80	1	YES
66	FEMALE	$55,204.70	1	YES

1. Show all the steps you would follow to determine the first attribute to split on for the data included in the table, using the information gain criterion. (10)
2. Suppose that we remove the "Mortgage" class label from the dataset. Show all the steps you would follow when applying the k-means cluster analysis algorithm to the data, with k = 3. (10)

Use the Customer.CSV file attached to this assignment to answer the following questions.

3. Apply the C4.5 (J48 in WEKA) decision tree algorithm to this data. Show your resultant confusion matrix and comment on the accuracy, precision, recall, sensitivity and specificity of the model you constructed. (10)
4. This dataset is imbalanced and the initial results are therefore poor. Explain how you would address this issue during preprocessing, training and model evaluation. (10)
5. Reapply the C4.5 decision tree algorithm to your modified data and determine whether your modifications have improved the performance. (10)
6. Ensembles of classifiers, or so-called meta-leaners, are often used in order to improve the accuracy of base learners such as decision trees. Explore whether applying a Boosting ensemble, such as AdaBoost, to this dataset improve the performance. Show your results and discuss your findings. (10)
7. Explain what a global outlier is and suggest an algorithm that you could use to identify a global outlier in the dataset. List one global outlier that you found. (10)