Homework 2: Data Preprocessing

Release Date: October 3, 2017

Due Date: October 12, 2017 (upon the start of class)

1. (**30 points**) Suppose the data for analysis includes an attribute *Age* and a binary class attribute (with values Y and N). The age values for the data records (in increasing order) and their corresponding class values are as follows.
   Age: 10, 13, 15, 16, 16, 19, 20, 20, 20, 22, 27, 27, 27, 27, 27, 34, 42, 42, 42, 45, 52, 60, 65, 70
   Class: Y, Y , Y , N , Y , Y , Y , Y , Y , N , Y , N , N , N , N , N , N , Y , N , N, N, N, N

   (a) Discretize the data into 8 intervals by equal-width (distance) binning. **(10 points)**
   (b) Discretize the data by equal-depth (frequency) binning using a bin depth of 3. **(10 points)**
   (c) Discretize the data into two intervals based on entropy-based binning (maximizing information gain), and show the conditional entropy H(class|Age) of the best split **(10 points)** Hint: You may implement a small program to find out the best split.

2. (**30 points**) Using the data for *Age* in Question 2, answer the following:
   (a) Use min-max normalization to transform the value 40 into the range [0.0, 1.0]. **(10 points)**
   (b) Use zero-mean normalization to transform the value 40. **(10 points)**
   (c) Use decimal scaling normalization to transform the value 40. **(10 points)**

3. Apply stratified random sampling on the following data to draw a sample of size 8. The two groups are: a) age<30 and b) age > 30. Note: The number of data points drawn from each group should be proportional to the size of each group. **(10 points)**
Age: 10, 13, 15, 16, 16, 19, 20, 20, 20, 22, 27, 27, 27, 27, 27, 34, 42, 42, 42, 45, 52, 60, 65, 70

4. Consistency is a popular feature subset evaluation measure. Two objects are considered matching objects if their values of all features (except the class) match. For *2* matching objects, *an* inconsistency occurs if their class values are different. For *n* matching objects, the number of inconsistencies (i.e., inconsistency counts) are determined by $n - \max\{m0, m1\}$, where $m0+m1=n$, and $m0$ and $m1$ are the number instances for class 0 and class 1 respectively. Consider the following data with four binary features (F1, F2, F3, and F4) and binary class label C. Perform feature selection using *Sequential Backward Selection* (*SBS*) search coupled with *inconsistency counts* as the subset evaluation measure (smaller inconsistency is better). What will be the selected subset of two features (illustrate your solution by intermediate steps)? **(10 points)**

F1	F2	F3	F4	C
1	1	1	1	1
1	1	0	1	1
1	0	0	0	0
0	1	0	0	0
1	0	1	0	0
0	0	1	1	1
0	0	0	0	1
0	1	1	1	0

5. (1**0 points**) For the same data set above, if *Sequential forward Selection* (*SFS*) is used with the same subset evaluation measure, what will be the selected subset of two features (illustrate your solution)?

6. **(10 points)** Discuss which method, SBS or SFS, is better for the above data and why?