Machine Learning Assignment 1

M13471127 - Devarashetti, Akhil Kanna
M13500936 - Ghotra, Sandeep Singh
9/21/2019

Problem 1: Discretization

1. **Rounding off**: The code for rounding method will be found in “discretization1.m” file.
2. **CACC**: The algorithm of Class-Attribute Contingency Coefficient is taken from the MATLAB code that was provided with the assignment. The code for this problem will be found in “discretization2.m” file.

Problem 2: ID3

This ID3 algorithm has been written by us without using any libraries. The train vs test spit ratio is 70:30. The accuracy is measured as the percentage of datapoints classified correctly.

Tree Representation: The decision tree is represented with 4 arrays namely *node_names, node_values, edges* and *parents*. The *parents(i)* contains the index of the parent node of *nodes_names(i).* The *edges(i)* contains the edge value between *node_names(i)* and its parent node.

1. **Rounding off**: The code will be found in the file named “ID3_1.m”.
    a. **Results**: The algorithm produces 15-16 nodes in the tree for the discrete valued iris dataset. The accuracy is found to be between 75% to 93% in the 5 observations with different train and test datasets.
2. **CACC**: The code will be found in the file named “ID3_2.m”.
    a. **Results**: The algorithm produces 11-15 nodes in the tree for the discrete valued iris dataset. The accuracy is found to be between 91% to 96% in the 5 observations with different train and test datasets.

Comparing the results: The CACC algorithm takes longer to discretize the real values, but it produces a desirable decision tree with fewer number of nodes and higher accuracy compared to the rounding off discretization method.