Data Structure

Homework: Linked List and Decision Tree

March 31, 2023.

“Diagnosis_7features.csv” comprises a set of anonymous clinical data for 640 subjects. In this data set, each subject has one primary key, seven clinical data and a class label (1: positive, 0:negative). The ultimate goal of this data set is to develop a binary decision tree to classify each subject into positive case or negative case.

Construction of a binary decision tree

1. Use Gini index as the cost function for model construction.
2. Use linked list to construct a binary decision tree, i.e., each node with two child nodes.
3. To avoid overfitting, each leaf node is supposed to have at least 5 subjects.
4. For a leaf node, suppose the numbers of positive and negative cases are n1 and n2, respectively. If n1 >= n2, this node is categorized as a positive node. Otherwise, it is categorized as a negative node. The accuracy of your binary decision tree is defined as (the sum of n1’s in all positive nodes + the sum of n2’s in negative nodes)/640.

Note that this data set will be used for several topics in this course, including, searching, sorting, tree construction, and, if possible, classification. To make you programs expandable, you are required to design your program in an object-oriented fashion.

Notes:

1. Submit your source code so that the TA can run your program and reproduce your results.
2. Report your decision tree, including
    a. the entire tree structure,
    b. the feature and threshold value used in each node of the tree,
    c. the Gini index at each node
    d. n1, n2 and categorization (positive node or negative node) of each leaf node
    e. the accuracy of your binary decision tree