Assignment 1

CS 484

Decision Tree Learning

Submission Deadline: 01/28 11:59 pm

Problem 1: (15 marks)

Inspect the dataset titled lab01_dataset_1.csv which has a mixture of numerical and categorical data. Your task will be to write a function my_ID3( ) which can create a decision tree for the given dataset using the ID3 algorithm. However, before doing that, you will be have to perform some data processing tasks. Here are all the required tasks in order –

1. ID3 cannot handle continuous numerical data. Perform necessary operations to handle all continuous-valued attributes. Do not forget to show the output i.e., the updated dataset after handling continuous-valued attributes. (2 marks)
2. Next, you will have to ensure the newly obtained dataset is optimal and free of errors. Take appropriate actions based on the outcomes.
    a. Check if the dataset has any missing values. (1 mark)
    b. Check if the dataset has any redundant or repeated input sample. (1 mark)
    c. Check if the dataset has any contradicting <input, output> pairs. (1 mark)
3. Your function my_ID3( ) should operate in a manner such that after ever round of decision making, it will output the attributes and its associated gain, with a message stating "Attribute X with Gain = Y is chosen as the decision attribute". Once your function completes, it should output the decision tree. The representation of the decision tree is upto you. You can choose either a textual representation or a graphical one; either is fine. (10 marks)

This dataset is relatively small and easy to understand just by looking at it. But you must perform all the above tasks via coding. Brute-forcing the answers or directly solving the mathematics involved in ID3 without coding it in Python will not get you a score.

Problem 2: (10 marks)

Inspect the dataset titled lab01_dataset_2.csv which also has a mixture of numerical and categorical data. For this problem, you will use decision tree classifiers for supervised learning. In particular, you will be using the functionalities of the sklearn.tree library. The classification task using sklearn libraries work only on numerical-valued attributes, and not on categorical ones. (What to do now? Hint: Look up One-hot Encoding and Integer Encoding). Here are all the required tasks –

1. Restructure the dataset such that it has all numerical-valued attributes. (2 marks)
2. Perform supervised learning using decision tree classifiers. Employ the train-test split approach during the learning. (4 marks)
3. After the learning is complete, show the results by predicting the class of the test set. Display the results of the prediction and test set side-by-side. (2 marks)
4. Output the decision tree; it can be either a textual representation or a graphical representation. (2 marks)