Machine Learning (BITS C464) - Assignment 1

Submission Time & Date: 10AM, 28 June 2023

Maximum Points – 20

Project Description

The census-income dataset contains census information for 48,842 people. It has 14 attributes for each person (age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, and native-country) and a Boolean attribute class classifying the input of the person as belonging to one of two categories >50K, <=50K. The prediction problem here is to classify whether a person's salary is >50K or <= 50K given the attribute values. The data description, training data set and testing data set can be found in in Assignment 1 – Data.zip that is iploaded in CMS.

i. There are missing values for some of the attributes and data tuples. You should apply the appropriate techniques for handling missing values.

ii. There are some continuous attributes among 14 features. If required, you may have to discretize these attributes using most appropriate techniques.

iii. Construct the optimal sized decision tree for predicting whether the income of a given person is >50K or <= 50K using the census-income dataset from the US Census Bureau. The optimal sized decision tree can be obtained by the applying "Reduced Error Pruning" technique that you learned in class. A graph, number of vertices vs error, for training data, validation data and testing data is to be depicted. The validation data set should be taken to 50% of testing dataset and the remaining 50% should be used as testing data.

iv. Combine training and testing data points. Randomly select 67% of the data points as training data set and the remaining data points as testing data set. Build a decision tree using the same procedure as put up in (iii).

v. Compare the optimal decision tree in step (iii) and step (iv). Comment whether the two decision trees are the same or not and justify your observation.

vi. Interpretability: Write down the rules that could be derived from the decision tree. Comment whether these rules are intuitive or not.

vii. Construct Random Forest classifier and compare their results with the results of decision tress obtained in (iii) and (iv).

viii. Report: You should come up with a consolidated report on the methodology, techniques that are used in building decision trees. The detailed results along with appropriate graphs and tables should be included in the report. You should discuss strengths and weakness of the systems that have been built.