Machine Learning (BITS F464) - Assignment 2

Submission Time & Date: 10AM, 10 July 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 uploaded 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. Build a prediction model using the following classifiers to predict whether a person's income is <50K or ≥ 50K
    a. Naïve Bayes Classifier
    b. Logistic Regression
    c. Neural Network classifiers with 1, 2 and 3 hidden layers
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.
v. Put up a comparative study of the following algorithms and figure out the optimal classifier
    a. Naïve Bayes Classifier
    b. Logistic Regression
    c. Neural Network classifiers with 1, 2 and 3 hidden layers
    d. Decision Tree Classifier (optimal one you obtained in Assignment 1)
    e. Random Forest
vi. Report: You should come up with a consolidated report on the methodology, techniques that are used in building the above mentioned classifiers. The detailed results along with appropriate graphs (if required) and tables should be included in the report.