Project 3 - Adult Census Income

1 Overall

In this assignment, you are required to learn a model on an Adult Census Income dataset, and the goal is to predict whether income exceeds $50K/yr based on census data. There are 3 files storing training data (traindata.csv), training label (trainlabel.txt), and test data (testdata.csv), respectively.

2 Problem Specification

This data was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (Data Mining and Visualization, Silicon Graphics). The prediction task is to determine whether a person makes over $50K a year.

Each line in traindata.csv corresponds to a data sample in the training dataset and different features and each line in trainlabel.txt stores a label corresponding to the data sample in the traindata.csv. Each line in testdata.csv corresponds to a data sample. Features are defined as follows:

- age: the working age of each data sample, which is a numerical variable;
- workclass: type of work, where there are private, local government, etc., is a character type variable;
- fnlwgt: the number of observational representatives of a sample in a state;
- education: the level of education of each sample;
- education_num: the schooling year of each sample;
- marital_status: marital status of each sample;
- occupation: the occupation of each sample;
- relationship: the family relationship of each sample;
- race: the race of each sample;
- gender: the sex of each sample;
- capital_gain: a capital gain is a profit that results from a disposition of a capital asset, such as stock, bond or real estate, where the amount realised on the disposition exceeds the purchase price;
- capital_loss: capital loss is the difference between a lower selling price and a higher purchase price, resulting in a financial loss for each sample ;
- hours_per_week: sample weekly working hours;

- native_country: the country where the sample is from;
- income (trainlable.txt): income, where income is greater than 50K and less than or equal to 50K.

3 Requirement

Based on the training set, you need to train a model out of machine learning models (decision tree, nearest neighbor classifier, support vector machine, neural network as well as ensemble models and so on).

You need to predict the class labels of the test data points. The prediction corresponding to each of your model on the test dataset should be stored in a txt file. Each line of the file stores the predicted class label for the corresponding test data point. You need to submit a report to describe what models you used and additional operations you made.

In all, you need to submit the source code, your predictions (in a txt file), and the report (in a pdf file) before the deadline. Grading will be based on the testing accuracy, the writing of the report and the source code.

Report

You need to submit a formal report. As tips, some key points are as follows.

- What is the problem to solve?
- How do you preprocess the data?
- What are the models of classifiers you would like to try? Please discuss each model w.r.t. the problem.
- How do you evaluate the models?
- How do you conduct experiments?
- Please compare and discuss your results.
- What are the limitations and how would you address them in the future?
- Your report is NOT to simply answer these questions. You should properly organize them in a formal report.
- Language: English or 中文

Source Code

- Your code.
- A README file about how to set up the environment, and how to run your code.

Attention

1. How to submit: You should compress your report (in a pdf, “report.pdf”), source code (in a folder, “code”) and predictions (in a txt file, “testlabel.txt”) into one zip file. The zip file with “ID_name”, e.g., “10101010_张三”, should be submitted to Blackboard.
2. Plagiarism, 0%. You could discuss with your classmate about the mini project, but please remember no plagiarism. We will check your report and source code.
3. Score: 70 pts (report) + 20 pts (source code) + 10 pts (predictions)
4. Deadline: June 9, 23:55. No late submission