Session 39

Project III Question

1. Introduction

This assignment will help you to consolidate the concepts learnt in the session.

1. Problem Statement

This data was extracted from the census bureau database found at http://www.census.gov/ftp/pub/DES/www/welcome.html

Split into train-test using MLC++ GenCVFiles (2/3, 1/3 random).

48842 instances, mix of continuous and discrete (train=32561, test=16281)

45222 if instances with unknown values are removed (train=30162, test=15060)

Duplicate or conflicting instances : 6

Class probabilities for adult.all file

Probability for the label '>50K' : 23.93% / 24.78% (without unknowns)

Probability for the label '<=50K' : 76.07% / 75.22% (without unknowns)

Extraction was done by Barry Becker from the 1994 Census database. A set of

reasonably clean records was extracted using the following conditions:

((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0)) Prediction task is to

determine whether a person makes over 50K a year. Conversion of original data as

follows:

1. Discretized a gross income into two ranges with threshold 50,000.

2. Convert U.S. to US to avoid periods.

3. Convert Unknown to "?"

4. Run MLC++ GenCVFiles to generate data,test.

Description of fnlwgt (final weight)

The weights on the CPS files are controlled to independent estimates of the civilian

noninstitutional population of the US. These are prepared monthly for us by Population

Division here at the Census Bureau. We use 3 sets of controls.

These are:

1. A single cell estimate of the population 16+ for each state.

2. Controls for Hispanic Origin by age and sex.

3. Controls by Race, age and sex.

We use all three sets of controls in our weighting program and "rake" through them 6

times so that by the end we come back to all the controls we used.

The term estimate refers to population totals derived from CPS by creating "weighted

tallies" of any specified socio-economic characteristics of the population. People with

similar demographic characteristics should have similar weights. There is one important

caveat to remember about this statement. That is that since the CPS sample is a

collection of 51 state samples, each with its own probability of selection, the statement

only applies within state.

**Dataset Link**

<https://archive.ics.uci.edu/ml/machine-learning-databases/adult/>

**Problem 1:** Prediction task is to determine whether a person makes over 50K a year.

**Problem 2:** Which factors are important

**Problem 3:** Which algorithms are best for this dataset

**NOTE: The solution shared through Github should contain the source code used and**

**the screenshot of the output.**

1. Output

***Solution:***

**Please fine the solution in the attached HTML file, I am unable to upload it directly to Github link.**

**Kindly enable the word file before clicking on the object so that it opens in a browser.**

****