Data Set For Project - STAT5375

Pham Minh Chau
R11685670
chaupham@ttu.edu

Oct 22, 2020

1 Data Source

The data set is obtained from UCI Repository [1] (Fig. 1). It was extracted by Barry Becker from the 1994 Census database.

Index of /ml/machine-learning-databases/adult

- Parent Directory
- Index
- adult.data
- adult.names
- adult.test
- old.adult.names

Figure 1: The data on UCI Repository

According to the data descriptions in $adult.names$, it contains $48,842$ observations splitting randomly into train ($adult.data$), test ($adult.test$) with ratio $2:1$. There are 15 variables in the data set namely *age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, native-country, income* (Fig. 2).

age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income
23	Private	239322	HS-grad	9	Divorced	Sales	Not-in-family	White	Male	0	0	40	United-States	<=50K
66	?	160995	10th	6	Divorced	?	Not-in-family	White	Female	1086	0	20	United-States	<=50K
33	Private	202642	Bachelors	13	Separated	Prof-specialty	Other-relative	Black	Female	0	0	40	Jamaica	<=50K
29	Private	255817	5th-6th	3	Never-married	Other-service	Other-relative	White	Female	0	0	40	El-Salvador	<=50K
44	Self-emp-inc	357679	Bachelors	13	Married-civ-spouse	Farming-fishing	Husband	White	Male	15024	0	65	United-States	>50K

Figure 2: Sample of the data set from UCI Repository

In this project, I choose $adult.data$ as the full data set (ignore $adult.test$) and retain only nine variables instead of all of them. These variables are *age, fnlwgt, education-num, marital-status, sex, capital-gain, capital-loss, hours-per-week, income* (Fig. 3).

[1] http://archive.ics.uci.edu/ml/machine-learning-databases/adult/

2 Data Description

The full data set for the project is from file $adult.data$ in which contains $32,561$ observations (Fig. 3). There are nine variables (six quantitative and three categorical) as following:

- *age*: (quantitative) age of an individual
- *fnlwgt*: (quantitative) the number of people the census believes the entry represents [2]
- *education-num*: (quantitative) the highest level of education achieved in numerical form [2]
- *marital-status*: (categorical) Marital status of the individual. Seven possible values are *Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse*. Note that *Married-civ-spouse* corresponds to a civilian spouse while *Married-AF-spouse* is a spouse in the Armed Forces [2]
- *sex*: (categorical) biological sex of the individual (*Female*, *Male*)
- *capital-gain*: (quantitative) capital gains for the individual
- *capital-loss*: (quantitative) capital loss for the individual
- *hours-per-week*: (quantitative) the total number of hours the individual has reported to work per week
- *income*: (categorical) whether or not the annual income of the individual is more than $\$50,000$ ($<= 50K, > 50K$)

age	fnlwgt	education-num	marital-status	sex	capital-gain	capital-loss	hours-per-week	income
23	161708	10	Never-married	Female	0	0	20	<=50K
34	206297	13	Married-civ-spouse	Male	0	0	48	>50K
36	212143	13	Married-civ-spouse	Female	0	0	20	>50K
52	318975	9	Divorced	Female	0	0	40	<=50K
34	178615	9	Married-civ-spouse	Male	0	0	40	<=50K

Figure 3: Sample of the data set for the project

[2] https://rpubs.com/Net/IncomeLevelClassification