Big Data Analytics Project : Dataset Adult from UCI

Mohammed Meftah, Raunaq Paul and Binxiang Xiang

16^{th} January 2017

Contents

1 Introduction

For this project, we have decided to work on the dataset Adult from UCI. This is a classification dataset with two classes, as we have to understand and predict whether a person earns less or more than 50k dollars a year, according to the provided features. Section 2 describes in details the used dataset Adult. Section 3 highlights how we first processed data. Section 4 shows some statistical analysis on the dataset. Section 5 presents the methodology and the performances of different algorithms (list of tried algo to add) on the prediction of our target variable.

2 Dataset description

The raw dataset has 48842 rows and 15 columns including the target variable. This data is originally from the american census bureau and shoulb be representative of the american population. This dataset is mixing continuous variables and discrete variables as follows :

age : continuous, age.

fnlwgt : continuous, weights used to build the dataset, not useful for us.

workclass : discrete, 8 categories representing the status of the company people work for (Private, Federal-gov, Never-worket ...).

education : discrete, 16 categories representing the education background (Bachelors, Masters, Doctorate, Highschool grad ...).

education-num : continuous, number of years of education.

marital-status : discrete, 7 categories representing the marital status (Divorced, Never-married, Separated ...).

occupation : discrete, 14 categories representing the profession (Tech-support, Sales, Exec-managerial, Other-service ...).

3 Data processing

4 Analysis

5 Methodology and results

6 Conclusion