Handling Missing Value

Arnold Jiadong Yu

March 7, 2019

Before we handle the missing value, eduction-num attribute is removed since it is the same as education attribute.

The missing values are marked as ' ?' in the dataset. The two strategies are implemented for handling the missing value:

1. Since the missing values are only 7% of the dataset, we can remove all the rows that containing missing values. This won't create additional bias.

2. Since the missing values are all categorical data. We replace the missing value by the value which has most frequency in that column. It makes more sense for workclass, country attributes but not really for occupation. This will create additional bias.

We take the input dataset and output 4 modified dataset. They are

1. train_removeMissing.csv

2. test_removeMissing.csv

3. train_replaceMissing.csv

4. test_replaceMissing.csv

Instructions on compiling and running program:

IDE: Anaconda-Spyder with python 3.6

1. Open HandleMissingValue.py.
2. Set to current directory (where the HandleMissingValue.py is).
3. To generate file 1 and 2. Run line 1-52 in the code.
4. To generate file 3 and 4. Run line 1-37 and 54-70 in the code.