Anshul Vyas
Mini Project 2
Spring 2015

Strategies for handling missing values

Here we do data preprocessing also:

Remove FNLWGT and EDUCATION-NUM attributes as they are of little use to our analysis. EDUCATION-NUM is the duplicate of EDUCATION.

The missing values in the adult dataset are marked by '?'.We can implement these two strategies for handling the missing values in the adult dataset:

1. Deleting all the rows containing the missing value. This seems a far-fetched method but we only lose a couple thousand observations when we have a pool of more than forty thousand.

2. This could have also been done by replacing the missing value

   - By the most common value in that column - for categorical attributes
   - By averaging (taking the mean) - for continuous attributes.

We apply these two strategies to produce two sets of adult dataset. Namely – *dataset-removed-missing.csv* and *dataset-replaced-missing.csv*

Python-script: handle-missing.py

Usage: python handle-missing.py [INPUT-FILENAME]

Input: dataset-data+test.csv

adult.data.csv

adult.test.csv

Output: dataset-replaced-missing.csv

Dataset-removed-missing.csv

The input files are provided in the INPUT folder. The output files generated are there as an example.