CS395 – Final Project

Part 1

Date: Feb 27th, 2019

By: Lauren Simms and Joshua Eli Swick

1. Choose and describe your dataset. The dataset you choose should be publicly available. It helps if it is labeled so you can verify the metrics (e.g., accuracy).

   We chose a dataset that describes the income census data from 1994. It includes demographic information such as age, education, marital-status, sex etc. We will be able to do supervised classification on this dataset.

   https://archive.ics.uci.edu/ml/datasets/Census+Income

2. Describe your data. You will probably need to have it in a single file for processing, so if it is in multiple files now, you need to do this first.
   a. What type of data is it?

      Multivariate

   b. Is it nominal, ordinal, interval or ratio?

      Nominal categories and ordinal ranges.

   c. What are the ranges or possible values?

      Categorical text and continuous integers in the range of zero to hundreds of thousands.

3. Provide a few sample rows of the dataset.

age, workclass, fnlwgt, education, education-num, marital-status, occupation, relationship, race, sex, capital-gain, capital-loss, hours-per-week, native country, salary

19, Private, 168294, HS-grad, 9, Never-married, Craft-repair, Own-child, White, Male, 0, 0, 40, United-States, <=50K

54, ?, 180211, Some-college, 10, Married-civ-spouse, ?, Husband, Asian-Pac-Islander, Male, 0, 0, 60, South, >50K

39, Private, 367260, HS-grad, 9, Divorced, Exec-managerial, Not-in-family, White, Male, 0, 0, 80, United-States, <=50K

49, Private, 193366, HS-grad, 9, Married-civ-spouse, Craft-repair, Husband, White, Male, 0, 0, 40, United-States, <=50K

23, Local-gov, 190709, Assoc-acdm, 12, Never-married, Protective-serv, Not-in-family, White, Male, 0, 0, 52, United-States, <=50K

20, Private, 266015, Some-college, 10, Never-married, Sales, Own-child, Black, Male, 0, 0, 44, United-States, <=50K

45, Private, 386940, Bachelors, 13, Divorced, Exec-managerial, Own-child, White, Male, 0, 1408, 40, United-States, <=50K

4. Transforming the data.
    a. Describe any transformations (e.g., one hot encoding) or normalizations you plan to do to prepare the dataset.

    The text fields will need to be transformed into numerical values for categorical representation. The only data point that will need normalized is the fnlwgt (Final Weight) as the data point is measured in hundreds of thousands.

    a. How will you break the data into test and training sets? At what ratio?

    The total dataset is 48842 rows so we will likely use an 80:20 training to test ratio. The data will need a little cleaning as some values are missing. We do not expect the total number of instances be reduced by too much.

5. Describe what problem you are trying to solve in one paragraph.

Although the data set is quite dated, attempting to develop an accurate model for predicting income categories from census data will prove useful to the learning process. We hope that the exercise will act as a scaffold to more difficult and relevant models. More literally, we are attempting to predict an individual's salary category based on demographic data.