CS289 Initial Report: Methods of handling missing data for classification

Rafael Valle[†] Jason Poulos[‡]

November 2, 2015

1 Motivation

Methods of handling missing data for neural networks classification model

Given that we plan to use NNets for income prediction ($income \geq \$50K/yr$) on the Adult dataset, we must handle missing data. This is less problematic for ML models such as random forest, decision trees, etc.

Item nonresponse is a common problem in survey data in several domains. Several techniques for data imputation (replace missing values with plausible ones) and direct estimation(all missing data is analyzed using a maximum likelihood approach) have been developed **?**.

Proper statistical adjustement of missing data is very important, as naive solutions might introduce bias. We're interested in using the data to train neuronal networks. This must be taken into account becouse higher input values will result in a higher activation input.

In this project, we plan to evaluate different data imputation and direct estimation techniques within the context of using a neural network income classifier on the ADULT dataset. We plan to compare our results to previous techniques and models, such as bla bla bla, that addressed this same question.

2 Data

We plan to experiment with the Adult data set and then, given the results, move to the a larger sensus dataset.

[†]rafaelvalle@berkeley.com.
[‡]poulos@berkeley.edu.

2.1 Adult data set

Characteristics:	Multivariate	Observations:	48842	Area:	Social
Features:	Categorical, Integer	Number of features:	14	Date Donated:	1996-05-
Associated Tasks:	Classification	Missing Values?	Yes	Number of Web Hits:	559776

2.2 Benchmarks

2.3 Exploratory data analysis

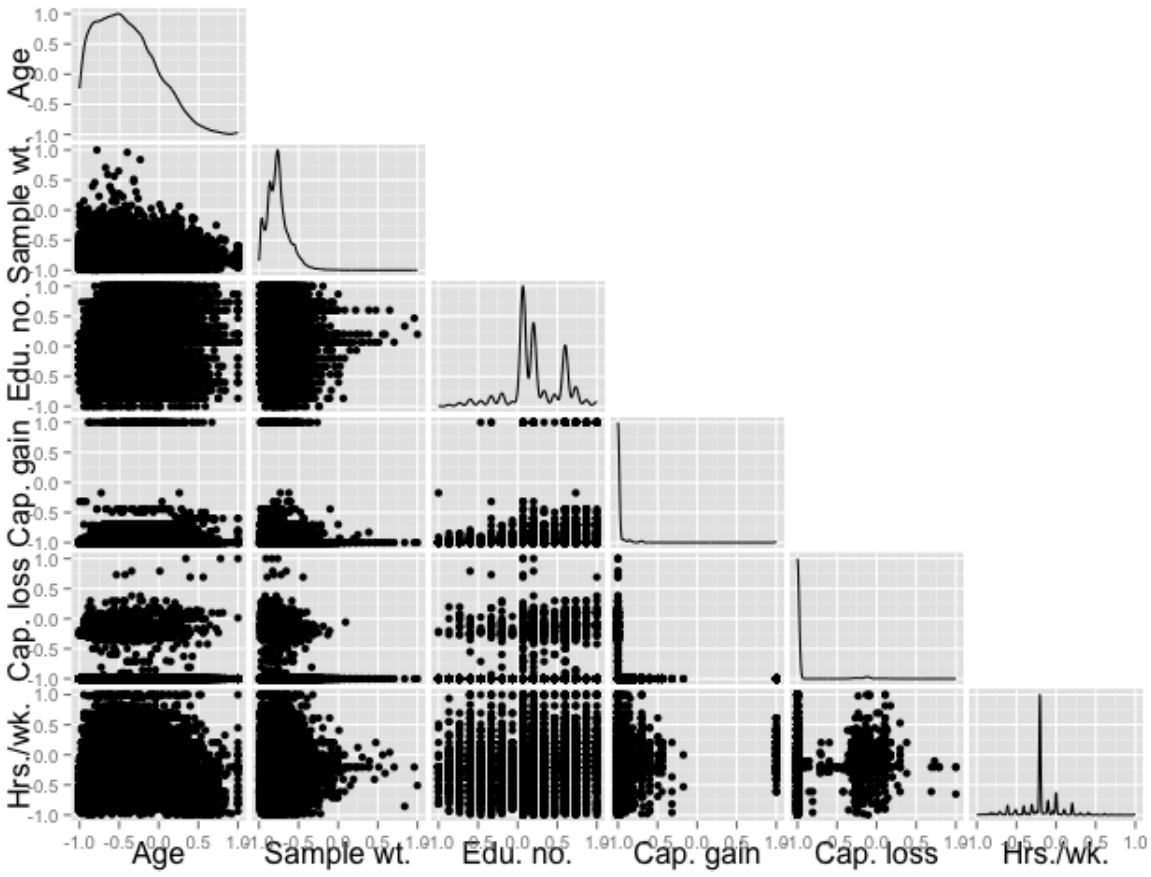

Figure 1: Pairwise correlations and distributions (diagonal) for all continuous features in Adult dataset. Feature values standardized to midrange 0 and range 2 (i.e., minimum -1 and maximum 1).

3 Methods

3.1 Techniques for handling missing data

Assuming that the techniques below are easy to implemet, we would like to compare their efficiency in imputing the missing values.

1. Basic Statistics : Replace the missing data with the mean or median of the feature vector. This is the most naive approach and since the missing variables in the ADULT dataset are all categorical, using the mean is not appropriated.

2. One-hot : Create an binary variable to indicate whether or not a specific feature is missing. This technique was mentioned by Isabelle

3. Nearest Neighbor Imputation : Recursively compute the K-Nearest Neighbors of the observation with missing data and assign median of the K-neighbors to the missing data. This technique is used in airbnb's fraud detection algorithm and explained in their website.

4. Logistic Regression : train a logistic regression model with all features except the feature with the missing variable to predict the missing value.

5. Bagging : Use one bag tree model for each predictor based on all other predictors.

6. Factor analysis : Perform some sort of factorization on the design matrix, project the design matrix onto the first two eigen vectors and replace the missing values by the values that might be given by the projected design matrix.

7. Find other features with distribution similar to the feature containing missing data and use this information (e.g. correlation) to fill in in the missing data. However, if two features are highly correlated, it might be better to remove one of them.

3.2 Neural networks for classification

4 Anticipated results